

**USE OF RELEVANCE CRITERIA ACROSS STAGES OF DOCUMENT
EVALUATION: A MICRO LEVEL AND MACRO LEVEL ANALYSIS**

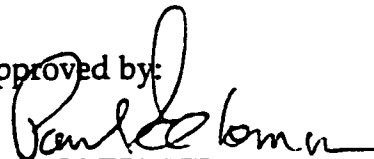
by
Rong Tang

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Ph.D. in the School of Information and Library Science.

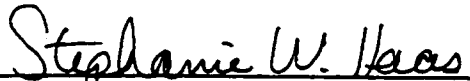
Chapel Hill

1999


Approved by:



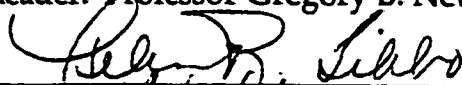
Advisor: Professor Paul Solomon



Reader: Professor Stephanie W. Haas



Reader: Professor Gregory B. Newby



Reader: Professor Helen R. Tibbo



Reader: Professor Jack L. Vêvea

UMI Number: 9954723

Copyright 1999 by
Tang, Rong

All rights reserved.

UMI[®]

UMI Microform 9954723

Copyright 2000 by Bell & Howell Information and Learning Company.

All rights reserved. This microform edition is protected against
unauthorized copying under Title 17, United States Code.

Bell & Howell Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

Chapter 1

INTRODUCTION

Introduction

It is widely accepted and well evidenced in the research literature that people's judgments of relevance change over the course of information seeking as they interact with the materials that they find. As people perform searches on online retrieval systems, their decisions regarding the relevance of a given document, based on the information available in a bibliographic surrogate, is not always consistent with the relevance evaluation that they make after reading the actual document. It is quite normal that an item initially estimated as relevant becomes only marginally useful in the end, and some items previously perceived as highly relevant even turn out to be completely irrelevant after the documents are sought and examined in full.

There are a variety of factors, as suggested by scholars, that contribute to the dynamics of relevance judgments. One apparent factor is the change in the formats of the materials being evaluated. Judgments based on a bibliographic record may differ from the judgments based on a full-text, due to the differences between both

the quantity and quality of the information contained in a surrogate and in a full-text. Besides the textual differences between formats of documents, change in relevance judgments is also perceived by researchers (Schamber, Eisenberg, & Nilan, 1990; Wilson, 1973; Harter, 1992; Wang, 1994) as originating from two important factors: change in judgment situations and change in users' cognitive states.

Users' relevance evaluation is sensitive to the situation of judgments. People's judgments during the process of information seeking and document evaluation are situated in a particular moment in the process, and, thus, their perceptions of the relevance for a given document change corresponding to the variations in their environment. Relevance judgments are also dependent on users' knowledge states. As people read through a set of materials, they obtain a better understanding of the needed information and the tasks at hand. This advancement in cognitive state, hence, is a third salient and legitimate reason for the dynamics in relevance evaluation.

The dynamic nature of relevance judgments can be further explored by examining the criteria that people employ to determine the relevance of documents at various stages of their research process. Since users' relevance judgments are dependent on a) the format of documents, b) their cognitive state, and c) the situation of information use, it appears that users apply different criteria to assess the relevance of a document at different stages of information seeking and document selection.

Two recent studies investigated the use of relevance criteria at different time points of document selection. While White and Wang (1997) found a number of new criteria both at the stage of reading (scan or read the actual documents) and at the stage of citing (cite the documents in the written product) in comparison with the initial stage of selecting (based on the evaluation of bibliographic surrogates), Bateman (1998a, 1998b) did not find statistically significant across-stage differences in participants' use of criteria for highly relevant documents. Both studies are exploratory in nature; nonetheless they provide good conceptual frameworks and empirical protocols for studying change in the use of relevance criteria.

The goal of the dissertation research is to investigate the dynamic process of relevance judgments by examining the specific criteria that people employ to select documents to meet their information requirements as they move through different stages of a document selection process.

Context of the Study

Research on relevance judgments has developed in the context of information retrieval. The notion of relevance was initially viewed as an objectified constant embedded in the mechanism of document retrieval, functioning as a measurement criterion for system performance. The advancement of information retrieval (IR) research saw the emergence of a variety of user-oriented paradigms, contrasting with the conventional system-oriented approach. Along with the evolution and expansion in the conceptual frameworks of IR, empirical research on relevance and relevance judgments began to develop its own sets of variables and units of

analysis. As a result, studies on relevance judgments have matured into an independent realm of research. Some scholars, represented by Hert (1997), express serious concern about the fact that research on relevance judgments has separated itself from its general context of IR and has grown to be perhaps distant from the consideration of the IR process as a whole. Even though the findings of such research provide some insights for system design, Hert asserts that only in the context of providing design advice for IR systems will the study of relevance judgments truly find its value and utility.

IR research has undergone a number of distinct phases in revolutionizing its theoretical framework. The general trend of IR research manifests a conceptual evolution moving from the system side of the IR equation to the human side of the equation. The conventional IR research employs a system-based paradigm, with a focus on maximizing system performance by improving retrieval algorithms. The user-based paradigm emerged as an alternative framework, and this framework promotes the role of the user in the IR process. The user-based paradigm views people as a central element of the retrieval process, and places attention on understanding information needs and describing behaviors in interaction with IR systems. The user orientation has gradually led to a third paradigm of IR, the cognitive paradigm. This paradigm zooms in to focus on the cognitive dimension of the IR process. The cognitive paradigm argues that the essence of the IR process is the interplay among various cognitive elements and structures. The quality of information retrieval lies in accelerating people's cognitive involvement and improving their cognitive states (Ingwerson, 1992, 1996). The fourth paradigm for

IR research, as suggested by Hert (1997), is a process-oriented paradigm. Hert (1997) believes that the process-oriented paradigm combines the virtues of all previously mentioned IR paradigms, with a focus on the dynamic nature of the IR interaction, as specified by stages of progression in time-space of the information retrieval process.

Following the conceptual movements under the general IR umbrella, the theory of relevance has had its share of adjustments and modifications. Specifically, the notion of relevance has evolved from a traditional system-based concept to a user-defined concept, then to a cognitive approach of relevance, and finally to a dynamic or process-oriented view of relevance. The focus on the IR system considers relevance as an objective topical relationship between a query and a document. User-defined relevance claims that the study of relevance should be based on end-users who have actual information needs. The cognitive approach further proposes that relevance is essentially an effect resulting from the cognitive improvement of users as they evaluate retrieved documents. The dynamic or process view of relevance emphasizes the dynamic, situational aspects of relevance judgments. This last school of thought is represented by a series of studies that describe situated change in relevance judgments during the process of document selection.

Relevance judgments are a very important part in the setup of an electronic bibliographic system. As Tibbo (1993) noted, the mechanism of current library services, specifically online retrieval, makes relevance judgments a necessary and almost inevitable step in an end-user's searching for information. Online retrieval

databases contain collections of bibliographic records. A bibliographic record serves as a surrogate for an actual document. It contains citation information, indexing information, and in most cases, an abstract of a document. To locate documents that are useful to an information need, someone normally would first examine a list of surrogates displayed by online databases to select a few items, then go to find full-text documents based on the selection. The existence and structure of these bibliographic databases transforms document searching into a process that consists of a series of evaluation stages with relevance judgments on document surrogates coming in as an early stage of the process. This model continues to apply in searching the World Wide Web.

In building document retrieval systems or bibliographic databases, abstracting and indexing services (A & I services) play a fundamental role. However, A & I services normally require tremendous amounts of time and human resources. With developments of technology and computer engineering, it is possible to implement bibliographic systems that mount full-text files. In current practice, there are several databases, for example, Lexis/Nexis, Westlaw, and Dialog, which offer at least some full-text documents. This has resulted in a debate as to whether it is still economically feasible to maintain bibliographic systems that retrieve only text surrogates.

Several studies have investigated the retrieval performance of full-text systems, and the results are mixed and contradictory. For instance, Blair and Maron (1985) studied the STAIRS system, and found low recall by the system. They, therefore, concluded that full-text retrieval performs poorly with respect to mission

critical tasks. Tenopir's (1985) study had different results. She found that full-text retrieval performed well in comparison to bibliographic systems of various kinds (abstract, controlled vocabulary, bibliographic union). However, Blair (1996) questions the validity of Tenopir's results and provides further arguments that challenge the capability of full-text retrieval. To date, researchers still have not yet reached agreement on the virtues of full-text retrieval, and no one is certain about whether such a format should replace other kinds of retrieval systems. Consequently, while full-text systems coexist with surrogate-based bibliographic databases, the latter remain the mainstream system in the overall scene of bibliographic retrieval.

Statement of the Problem

With the research on relevance judgments moving towards a process-oriented framework, and with IR practice still based on surrogate retrieval while full-text retrieval is becoming more and more available, it seems important to investigate and compare people's relevance judgments across different stages of document selection. This research aims to investigate users' relevance judgments from the perspective of the actual employment of relevance criteria at two distinct stages of the document selection process.

A document selection process begins to unfold when a person has a need to search for some literature for the purpose of completing a task, such as composing certain form of scholarly product on a subject matter. Typically, a document selection process involves people searching documents through online databases,

making judgments about the relevance of the retrieved items based on the information contained in the bibliographic records, and then proceeding to collect the actual documents and reading full-texts to determine how useful the items are in helping them to accomplish their tasks. This research focuses on two stages in the process: evaluation of bibliographic records and evaluation of full-text documents.

Specifically, the research addresses the following questions:

- During the process of document selection, what makes users consider a document to be relevant?
- What criteria do users employ at the stage of record evaluation and what criteria do they use at the stage of full-text evaluation?
- Are criteria used for an early evaluation period consistent with the ones used for a late evaluation period?
- What criteria are important for evaluating bibliographic records and what criteria are important for evaluating full-text documents?
- What criteria are used most frequently at each of the two stages?

Purpose of the Study

The purpose of the research is to assess the use of relevance criteria at the stage when a bibliographic record is evaluated (Stage 1) and at the stage when a full-text article is read (Stage 2). Comparisons of the use of criteria between the two stages describe change or evolution in people's use of relevance criteria as they progress from one stage of the document selection process to another.

Nature of the Study

The issue of the use of relevance criteria in the process of document selection involves multiple factors and is both complex and subtle. The investigation of this issue cannot be well rounded unless the problem is approached through multiple points of view. This research was conducted under the philosophy of methodological pluralism, which promotes the use of multiple methods in the investigation of a research question. The findings were based on the independent results of a laboratory experiment and a naturalistic study. I believe that methodological pluralism not only is appropriate to investigating the specific research problems already mentioned, but also to studying other issues related to relevance or information retrieval in context.

Significance of the Study

The significance of the research is two-fold: the study contributes to the conceptual advancement of theories of relevance and document evaluation, and the study suggests design possibilities for bibliographic retrieval systems.

By investigating the criteria used at the two stages of document evaluation, the study enriches the understanding on the human behavior in making relevance judgments. As relevance is a multidimensional concept, its true nature will only become evident by mapping and examining real-life document selection processes. The study collects data on various classes of relevance criteria; it therefore, extends our knowledge about specific aspects of relevance and explores the possible relationships and interactions among various dimensions of relevance. Since the study focused on the two consecutive stages of document evaluation, it, in particular, increases our knowledge of the dynamic nature of document evaluation with people's judgments situated in critical moments in a continuing natural process.

The results of the study enhance and enable reconstruction of previous theories on criteria for relevance judgment. Over the years, empirical investigations of criteria have concentrated on eliciting relevance criteria from users of IR systems. This line of research has accumulated a rich, yet overwhelming number of criteria. It is difficult to reach a consensus in classifying these criteria because many of them have multiple meanings and are situationally dependent. Moreover, it is generally assumed in the literature that these criteria comprise all factors of relevance and hence are applicable to relevance judgments in general. Seldom has any study

attended to the fact that criteria that are used for judging the relevance of document surrogates could differ from the judgments based on full-text documents. This research examines criteria for specific stages of document evaluation and investigates the corresponding changes in the use of criteria. As a result, it deepens our knowledge about the use of relevance criteria on the basis of the specific stages involved.

The findings of the study provide concrete ideas for the design of bibliographic retrieval systems. For example, based on the in depth mapping of use of criteria and change from Stage 1 to Stage 2, I offer a design concept for an interactive retrieval system that allows users to specify the criteria that they would use for evaluating documents and indicate their preferences for criteria. Besides this specific idea concerning interactive feedback retrieval, there are many other issues emerged from the data analysis of the research that would also be valuable for improving bibliographic retrieval systems, including both surrogate-based retrieval systems and full-text retrieval systems.

Chapter 2

RELATED LITERATURE

Introduction

The dissertation research builds on a great quantity of literature pertaining to the topics of relevance and end-users' relevance judgments. This review of related literature opens with a synthesized overview of theories of relevance, and follows with a summary of empirical studies of criteria for relevance. The third category of the literature covered deals with the issue of relevance judgments based on different formats of documents, and the last section includes research work that identifies and classifies the decision stages in the process of document selection.

Theories of Relevance

Theorists from different disciplines presented a variety of definitions of relevance. In the context of human perception, relevance is viewed as concerning a piece of information that contributes to confirming or rejecting a hypothesis describing the state of affairs of environment. Bruner (1973) suggests that in the actual process of perceiving a physical stimulus, "relevant information, or a relevant cue, refers to stimulus input which can be used by the subject for confirming or

infirming an expectancy about the environment" (p. 98). In the fields of communication and human cognition, relevance is perceived as determined by cognitive improvement resulting from the processing of information. Sperber and Wilson (1995) point out that "an assumption is relevant in a context if and only if it has some contextual effect in that context" (p. 122). Such a contextual effect, as described by Sperber and Wilson, is mainly signaled by the cognitive movements of an individual as the interaction with communicated information brings about some advancement in this person's cognitive state.

In the arena of information retrieval, relevance is primarily thought of as a relation between two or more entities involved in the process of retrieval. Saracevic (1970) presented an algorithmic definition of relevance which "displays certain relationships among parts while permitting a rather free manipulation of the parts" (p. 205). According to Saracevic, "Relevance is the A of a B existing between a C and a D as determined by an E" (p. 47). In this algorithm, A is the "gauge of relevance," B denotes the "aspect of relevance," C refers to the "object judged," D serves as the "frame of reference," and E is the "assessor." Table 2.1 displays the subelements under each facet of relevance.

Table 2.1
The Algorithmic Definition of relevance

A Gauge of Relevance	B Aspect of Relevance	C Object Judged	D Frame of Reference	E Assessor
Measure Degree Extent Quantity Dimension Judgment Estimate Appraisal Relation	Utility Importance Matching Informativeness Appropriateness Satisfaction Connection Fit Similarity Applicability Closeness Usefulness Bearing	document doc. representation references textual form fact info. provided	question question representation research stage information need point of view use of orientation treatment	requester intermediary expert librarian info. specialist delegate user person judge

(Source: Saracevic 1970, p. 48.)

Saracevic's definition of relevance demonstrates that the notion of relevance, in the context of retrieval, is a multi-faceted concept that carries much more rich and diverse connotations than it does when used in other fields. Recently, in a comprehensive review of the relevance literature, Mizzaro (1997) developed a simpler structure in defining relevance. He specifies that relevance is a relation between the entities of two groups: The first group contains either "Document," "Surrogate," or "Information," the second group includes either "Problem," "Information Need," or "Query." With this structure, relevance can be operationally defined either as a relation between a surrogate and a query, or a relation between a document and an information need, and so forth.

Because there are a variety of entities involved in the notion of relevance, researchers have projected different views on relevance. Some believe that relevance is objective, pertaining mainly to the topical relationship between a query

and a document; others argue that relevance is largely subjective, related to users' cognitive structure. Still others suggest that the dynamic nature and time dimension of the relevance judgments defines the essence of relevance. In the following paragraphs, I will review three major views of relevance, including *Objective and Topical View of Relevance*, *Subjective and Cognitive View of Relevance*, and *Dynamic and Situational View of Relevance*. Each of the three views of relevance is described in connection with one specific concept. The *Objective and Topical View of Relevance* is closely associated with the term *Aboutness*, the *Subjective and Cognitive View of Relevance* is frequently expressed by the idea of *New Information*, and lastly, the notion of *Usefulness* often accompanies the *Dynamic and Situational View of Relevance*. In the end, I propose that relevance is a multivariate construct and that a multidimensional concept of relevance should be established to incorporate a variety of aspects of relevance, thereby constituting an overall understanding of the concept.

Objective and Topical View of Relevance and Aboutness

Regarding the nature of relevance, a clear distinction was made in the beginning of IR research between a topical, objective view of relevance and a cognitive, subjective view of relevance. In an early work on information retrieval, Vickery (1958a, 1958b) points out that there are limits and levels of relevance that influence the design of retrieval systems. At a primary level, a retrieval system operates by distinguishing a collection of documents based on the principle of "literary warrant." According to Vickery (1958a), literary warrant means that "if a

given subject has appeared in the literature, and if it is desired to retrieve documents relevant to that subject, then it must be possible to represent the subject by the descriptors used in the system” (p. 863). In other words, the idea of literary warrant is to ensure that the system retrieve “items ‘relevant’ to a particular sought subject” (Vickery, 1958b, p. 1277). Vickery (1958a) further indicates that every retrieval system is built on this literary warrant mechanism, and such a mechanism operates on word matching between the subject terms in a query and the subject terms in a document. Beyond this basic retrieval principle, Vickery proposes that a second criterion of relevance, serving as the upper level retrieval principle, is “user relevance.” Vickery suggests that the relevance of a document depends essentially on the individual who is making the judgments. One searcher may decide a document to be relevant while the other may perceive it differently. At the level of “user relevance,” relevance is highly subjective and individual-specific.

Relevance to a subject is most commonly expressed in the literature of IR as the relevance of “topicality,” or simply, topical relevance. Cooper (1971) uses the term “Logical Relevance” to describe topical relevance, and contrasts it with the concept of “Utility.”

In approaching the question of relevance in an information retrieval context, it seems natural to make at the start a rough distinction between what has been called *logical relevance*, alias “topic-appropriateness,” which has to do with whether or not a piece of information is on a subject which has some topical bearing on the information need in question and *utility*, which has to do with the ultimate usefulness of the piece of information to the user. (p. 20)

In general, topical relevance states that the degree of relevance is mainly a reflection of how much the topical content of an information object (i.e., a document) bears on, or matches the content of an information request or a query. The topical relationship between a document and a query as perceived by Cooper is “logical,” and may be recognized or inferred from texts. When relevance is defined as being reflected purely by logical topicality, it is automatically assumed that relevance is both objective and constant. Vickery (1958a) states that under the literary warrant criterion, “it is quite justifiably assumed that discriminations which have been relevant to authors in the past will be, to a greater or lesser extent, relevant to readers in the future” (p. 864).

The fundamental belief of conventional IR theory, as reviewed by Swanson (1977, 1986), is that once the document and the information request are presented in a written form, they are objectified and become independent of their creators. Traditional IR holds that “once an information need is objectified as a written request, the possibility arises that such a request is logically related to some document ... that relationship is then a basis for saying that the document is objectively relevant to the request. It is relevant whether or not anyone notices that it is relevant Relevance in this sense, being a link between a written request and a document, belongs to the world of objective knowledge” (Swanson, 1986, p. 391). Swanson (1977) further indicates that objective relevance operates on a frame of reference – *Frame of Reference 2*. This frame of reference is oriented to evaluate the relevance of a given document based on whether the document is “on the same

topic” as the request. The relevance judgment is considered essentially as “a judgment by the requester that a given document does deal with the topic that he requested, that it fits the description of the topic given in his request” (p. 140). In summary, the objective and topical view of relevance sees relevance as the logical topical relationship between a query statement and a document, and it further implies that relevance is generally recognizable, consensusable, and belonging to public knowledge (Foskett, 1972).

The concept of topical relevance relates directly to the notion of “aboutness” of a document. In theories of text processing, scholars have proposed that there are different types of aboutness. A primary distinction is drawn between the concept of aboutness (what a document is about) and the concept of meaning (what a document means to an individual). At an operational level of information retrieval, Maron (1977) suggests there are three forms of aboutness: *objective about*, *subjective about*, and *retrieval about*. The *objective about* (O-about) is a behavioral concept of about, and it is “obtained by considering an external or observer’s point of view” (p. 41). The *subjective about*, on the other hand, is related to an individual’s inner experience of what a document is about as the person reads a document. Maron (1977) explains that an S-about (subjective about) is “a relationship between a document and the resulting inner experience of its readers. It is a psychological concept and like similar psychological concepts it is very complex and cannot be analyzed further in objective terms” (p. 41). The third concept of about, the *retrieval about*, refers to “the information searching behavior of a class of individuals” as

reflected by the retrieval results. Maron further indicates that the R-about (retrieval about) is also objective and behavioral oriented.

Fairthorne (1969) introduces a somewhat different perspective to the understanding of “aboutness.” He contends that there is a difference between *intentional aboutness*, which is the author’s views and intentions of what a document is about, and *extensional aboutness*, which is the document aboutness as reflected semantically by actual units and parts of the text. The distinction between extensional versus intentional aboutness, as interpreted by Beghol (1986), is conceptually equivalent to the distinction between “aboutness” and “meaning.” Beghol proposes that the notion “aboutness” denotes that “a document has an intrinsic subject, an ‘aboutness’, that is at least to some extent independent of the temporary usage to which an individual might put one or more of its meanings” (p. 85). In Beghol’s terms, “topicality” is the same as “aboutness.” She further contends that the topicality (or “aboutness”) of a document is objective, independent, and constant. Beghol’s proposition on nature of aboutness is clearly illuminated in the following statements:

...texts of all kinds have a relatively permanent aboutness, but a variable number of meaning(s). . . . a document may have only one aboutness, but an unlimited number of meanings, differing according to the exact use a particular person may find for the document’s aboutness at a certain time. Indeed, the same document can have different meanings for the same reader at different times, but the document, itself unchanging, is assumed to possess a fundamental aboutness. (p. 85)

Beghol separates the notion of aboutness from the actual use of aboutness; the former is seen as containing one permanent form, the latter, however, has an unlimited number of forms, it is individual specific and it changes over time. Beghol's claim that a document possesses a permanent, unchanging aboutness supports the idea that relevance, as a reflection of the topical relationship between a document and a query, would be objective and constant in nature. On the other hand, Beghol's belief that different meanings of a document can be derived from different readers at different times, may be extended to imply that users have different perceptions of the relevance of a document, and therefore relevance is subjective, cognitive, dynamic and situational.

The inadequacy of an objective topical relevance has been recognized repeatedly in the literature of information retrieval. Researchers such as Froehlich (1994) and Barry (1994), for example, indicate that the objective and topical view of relevance does not account for subjective and psychological aspects of the individual who is making the decisions. The following comment made by Park (1994), best illustrates the common criticism of an Objective and Topical View of Relevance:

Topical relevance is context-free and is based on fixed assumptions about the relationship between a topic of a document and a search question, ignoring an individual's particular context and state of need. It is a unidimensional view of users' information problems, disregarding the changing nature of the individual's information problem and its subsequent impact on the search. It fails to focus on the complexity of the individual's background and task situation. (p. 136)

Bert Boyce (1982) also asserts that topicality is an operationally necessary but insufficient condition for user-oriented information retrieval. He first points out that the relevance judgment made by a judge is different from that by a requester. A judge is an individual who makes relevance evaluation for someone other than himself; a requestor, on the other hand, is the user who has an original information need and makes judgments based on his or her real needs. Boyce (1982) argues that “a judge who is not the requestor must . . . make a judgment based on topicality” (p. 105). However, to satisfy a user’s need, topicality is “surely insufficient” and that “something else is required” (p. 105). Boyce contends that “it is quite clear that any satisfaction of user’s need will be highly subjective and dependent on the knowledge state of the requester. One can get intersubjective agreement on the topicality of a document, but hardly upon the degree of personal satisfaction its presentation engenders. This can only be a function of the knowledge state of the requestor as it varies over time” (p. 105). Evidently, Boyce believes that there are other important elements beyond the objective relevance and the topical relevance. He further specifies that these elements have to do with the users’ knowledge state and the situational dynamics of judgments.

Subjective and Cognitive View of Relevance and Newness

The second view of relevance promotes the dimensions of human subjectivity and cognition in the process of relevance decision making. Relevance is not seen as associated only to the topicality of a text. In the domain of discourse comprehension, van Dijk (1979) proposes that *contextual relevance* is as important as

textual relevance. *Textual relevance* is reflected by specific textual structures as denoting the topical or thematical relevance of the text. *Contextual relevance*, as described by van Dijk, “is the assignment of a relevance value on the basis of any kind of contextual criterion, such as the interest, attention, knowledge, wishes, etc., of the reader” (p. 113). van Dijk suggests that relevance is contextually determined in that “the cognitive (and social, communicative) context defines what elements of a text are found important by a reader” (p. 119). Here, van Dijk adds the dimension of reader’s cognition to the understanding of relevance.

In the domain of information retrieval, the notion of relevance has also been perceived as containing cognitive and contextual elements. Foskett promotes the concept of “pertinence” in contrast with the concept of relevance. Foskett (1970) defines relevance as any piece of knowledge that can “fit in with the general pattern of a larger area ... it is a recognizable and recognized part of the consensus among experts in that area” (p. 91). He then states that “‘pertinence’ means that it fits in with the particular pattern that one individual is trying to construct in his own mind” (p. 91). Foskett claims that relevance is something that is objective and publicly agreeable, whereas pertinence is something that is subjective, and it is linked to an individual’s private knowledge and cognitive state. In a follow-up paper, Foskett (1972) reiterates the conceptual distinction between a public, objective relevance and a private, subjective pertinence. He states that while relevance is “belonging to the field/subject/universe of discourse delimited by the terms of the request, as established by the consensus of workers in that field,”

pertinence means “adding new information to the store already in the mind of the user, which is useful to him in the work that prompted the request” (p. 77).

Foskett’s concept of pertinence contains elements of individual subjectivity and cognition, which, are also the basis of Swanson’s notion of “Subjective Relevance.” Swanson (1977) indicates that contrary to general assumptions of Objective Relevance, Subjective Relevance looks at the relevance judgment as a subjective, evolving mental process on the part of the user. Swanson suggests that Subjective Relevance applies *Frame of Reference 1*, which is a frame of reference that treats relevance as a “creative, subjective mental act by the requester, expressing whether a document fulfills whatever information need prompted his request” (p. 129). This frame of reference also views relevance judgment as a process of “creation of new knowledge . . . The ‘relevance’ of a document is here taken to be a piece of new knowledge constructed by the requester in the light of some information need or deficit” (p. 139). Subjective Relevance implies that relevance does not exist prior to the judgment process, it is not “a property of a document and a request” (p. 139), and it is only formed during the judgment process when the new knowledge is generated.

The proposition that relevance is determined subjectively and cognitively is restated strongly in Harter’s theory of “psychological relevance.” Harter (1992) extends Sperber and Wilson’s claim that relevance is signaled by the cognitive improvement of an individual to the context of document retrieval and proposes that relevance is reflected solely by cognitive changes in the mental state of the user.

A document may be “on the topic” but may not necessarily initiate any cognitive changes in the users’ mental state. In such a case, Harter states, the document should not be considered as relevant. Harter puts a strong emphasis on the cognitive nature of relevance. However, his psychological definition of relevance appears to be too restrictive since it goes to an extreme to eliminate the considerations of other dimensions of relevance such as topicality and situationality.

A most intriguing construct repeatedly seen in the cognitive and subjective theories of relevance, is the addition of the term of “new” as in the phrases “new information” or “new knowledge.” Foskett describes “pertinence” as “adding new information to the store already in the mind of the user,” and Swanson believes the essence of “Subjective Relevance” lies in the “creation of new knowledge.” The variable “newness” seems to be strongly associated with subjective and cognitive view of relevance, and it therefore deserves further discussion.

In the context of discourse analysis and textual linguistics, the term “new information” is often contrasted with the term “given information.” A text is seen as embodying both types of information. Some information is given, which the speaker or writer assumes the audience knows of already, yet other parts are “new” elements that convey information not previously known. Hutchin (1978) describes a typical reader’s expectation as he or she approaches a document:

He comes to the document with an interest, a desire or a need to “improve” in some way his present state of knowledge. What he wants is a document which contains information that is “new” to him and which assumes no more knowledge than he has already. These, then, may be regarded as the basic conditions which must be satisfied: (i) the information conveyed as “New” (i.e., not presupposed) in the document must include some that the reader did not know before; and

(ii) the knowledge taken as “given” (i.e., presupposed) must be at a level lower than, or roughly equal to, that of the reader. (p. 177)

Many text linguists, represented by Robert de Beaugrande, also advocate that text processing is a dynamic process during which the text users' knowledge state, cognitive state, and expectations experience constant change and reconfiguration. Theorists further suggest that the most efficient text processing involves text objects containing information that provides a certain quantity of newness to the reader so that the reader would find the text to be interesting, intellectually compelling, and useful. Beaugrande (1980) suggests that the defining characteristic of a piece of information being informative is that it contains something new. He indicates that “the term INFORMATION can be taken to designate not the knowledge that provides the content of communication but rather the aspect of newness or variability that knowledge has in some context” (p. 103). In Beaugrande and Dressler (1981)'s seven standards of textuality, “informativity” is viewed as one of the principles that describes “the extent to which the occurrences of the presented text are expected vs. unexpected or known vs. unknown/uncertain” (p. 8-9). To further explicate informativity, Beaugrande (1980) suggests that there are three orders of informativity, which describes how much of the content of a text is new. The *first-order of informativity* holds the lowest level of newness, and therefore it has the greatest ease of processing and requires the lowest degree of cognitive involvement. The *third-order of informativity* contains “unusual and extremely interesting occurrences and is correspondingly hard to understand

and control” (p. 105-107). Despite the highest degree of interestingness, *third-order informativity* could result in serious problems during text communication in that it may create major discontinuities, gaps, and discrepancies in users’ minds. The *second-order of informativity* generates a nice balance between efficiency and effectiveness in text processing. This balance is reached when new information provides the reader with a healthy degree of cognitive involvement. Beaugrande believes that the *second-order of informativity* presents a good deal of new information that is cognitively acceptable to the readers without causing difficulties and breakdowns in understanding.

In Sperber and Wilson’s definitions of relevance, the cognitive improvement of an individual is perceived essentially as a process whereby the interaction of the new information and the old information give rise to further new information in the mind of the receiver. Sperber and Wilson (1995) point out “Some information is old . . . Other information is not only new but entirely unconnected with anything in the individual’s representation of the world. . . Still other information is new but connected with old information. When these interconnected new and old items of information are used together as premises in an inference process, further new information can be derived. . . When the processing of new information gives rise to such a multiplication effect, we call it relevant” (p. 48). Relevance thus functions as a contextual effect, which is achieved only when the addition of new information in connection with the old information, produces or derives further information.

As it is shown from the quotes earlier, the concept of newness is often defined in partnership with the word “understanding.” The new information included in a document must consider the user’s ability to understand. Boyce (1982) states that the purpose of a retrieval system is to be able to select a set of documents that is informative. He suggests that “to be informative a document must have at least two characteristics beyond topicality. It must be both understandable and novel” (p. 106). Here the novelty of information is put together with users’ understandability, and the two constructs are preconditioned on each other. The element of novelty corresponds to the newness of information, while the principle of understandability suggests that newness has to be at such a degree that it does not presume a knowledge level that is higher than the current knowledge state of the receiver. Boyce proposes that “if a document is both novel and understandable it will result in a transformation of the requestor’s knowledge state” (p. 106).

Dynamic and Situational View of Relevance and Usefulness

The idea of informativity and newness draws attention to the cognitive aspect of the individual who is making the judgments. There are some other definitions of relevance that also center on the user, yet from slightly different perspectives. For instance, *Situational Relevance*, proposed by Patrick Wilson (1973), suggests that relevance is “relevance to a particular individual’s situation” (p. 460). Wilson believes that the study of relevance judgments should investigate “the actual uses and actual effects of information: how do people use information, how their

views actually change or fail to change consequent on the receipt of information” (p. 458). Such a research orientation should provide “a complete description of the world as he [the user] sees it” (p. 460). From the situational view, relevance is seen as dynamic, and relevance judgment is perceived as evolving along with the change in time and the situation of judgment.

In 1990, Schamber, Eisenberg and Nilan proposed that there is a need to reexamine the concept of relevance, and that we ought to establish a dynamic, situational definition of relevance. Schamber et al. further outline the philosophical underpinnings of the dynamic, situational approach.

The dynamic, situational approach we suggest views the user – regardless of system – as the central and active determinant of the dimensions of relevance. We believe that relevance is a multidimensional concept; that it is dependent on both internal (cognitive) and external (situational) factors; that it is based on a dynamic human judgment process; and that it is a complex but systematic and measurable phenomenon. (p. 755)

Here a multidimensional concept of relevance is seen as containing both internal cognitive elements and external situational factors. The subjective and cognitive view of relevance fits well with the dynamic and situational view of relevance, since both views complement each other by focusing on different aspects of the end-users who make the relevance judgments.

In the study of relevance, the terms “usefulness” or “utility” often occur simultaneously with the term “relevance.” Cuadra and Katter (1967) suggest that *usefulness* refers to the use orientation of the information, in other words, *usefulness* is reflected by the intended use of the documents. Cooper (1973) considers *utility* to

be an antithesis to *logical relevance*. He defines “utility” as “a catch-all concept involving not only topic-relatedness but also quality, novelty, importance, credibility, and many other things” (p. 92). He further argues that the personal utility of a retrieval system’s output to users can be quantified and operationalized, serving as a measure of retrieval effectiveness.

The concept of utility, as in Saracevic’s description of the pragmatic view of relevance, relates directly to the cost-benefit aspect of the information being used. According to Saracevic (1975), the pragmatic view of relevance claims that it is not good enough for IR systems to provide relevant information, and that the true role of IR systems “is to provide information that has utility--information that helps to directly resolve given problems, that directly bears on given actions, and/or that directly fits into given concerns and interests . . . immediate pragmatism of information in one form or another is the ultimate, definite, final criterion” (p. 334). Saracevic further comments that such a pragmatic view reflects a limited, narrow interpretation of relevance by neglecting many other facets of relevance.

A comprehensive empirical investigation of the functionality of usefulness was conducted by Rees and Schultz (1967). In their design, the investigators intentionally contrast the concept of usefulness with the concept of relevance, the latter is defined to the participants as mainly concerned with the topical relatedness: “by relevance we mean the degree to which the document bears on, or has application to, the research you have heard described.” (p. 84). The authors point out that in making an evaluation of the relevance of a given document, the judge

should view relevance as only concerning the relationship between the document and the research, and nothing more. The notion of usefulness, on the other hand, represents a relationship among three entities, the document, the research, and the judge. Usefulness was defined to the study participants as “the degree to which the document would be useful to you as an individual. In other words, usefulness should take into account your interests, knowledge, experience, etc. in doing this research” (p. 85). Through analyzing judges’ self-reported criteria for assessing relevance/usefulness dichotomy, Rees and Schultz also found that in judges’ minds, the concept of usefulness not only includes *topical* relevance, but also reflects other personal utility aspects such as *newness* of the information (criterion 6a and 6b), *understanding* of the information (criterion 8a and 8b), and the *use* of information in relating to the actual *citing* (criterion 7).

A Multidimensional Cognitive Concept of Relevance

At the end of their paper, Schamber, Eisenberg and Nilan (1990) drew three conclusions regarding the nature of relevance and its role in information behavior:

1. Relevance is a multidimensional cognitive concept whose meaning is largely dependent on users’ perceptions of information and their own information need situations.
2. Relevance is a dynamic concept that depends on users’ judgments of the quality of the relationship between information and information need at a certain point in time.
3. Relevance is a complex but systematic and measurable concept if approached conceptually and operationally from the user’s perspective. (p. 774)

The idea of a multidimensional cognitive concept of relevance has been frequently referred to and generally praised in the relevance literature. However, in

actual practice, scholars have conceptually contrasted objective topical relevance from subjective cognitive and dynamic situational relevance, and viewed them as two confronting paradigms that are independent. As pointed out by Saracevic, the user-oriented researchers, who advocate the emphasis on the psychological, cognitive and situational aspects of relevance judgments, often appear to be overly critical of system-oriented scholars whose research model is based primarily on objective topical relevance. Saracevic (1996b) indicates that either approach has blind sides, and that concentrating on attacking each other's theoretical stands only worsens the bipolar isolation. Saracevic further claims that by insisting on the conceptual splits between the two camps, "we imposed on ourselves a limit of reductionism in approaching this problem. We approach it from either the system side or user side, in either case, this is limiting, reducing the problem to seeing and working only on one component of a complex problem" (p.23). He encourages an interaction between the two camps: "it is not one camp against the other but how can we incorporate the best features of both approaches and make them work jointly" (p. 22).

One conceptual extension from Schamber et al. (1990) and Saracevic (1996b)'s work is that the subjective and cognitive and the dynamic and situational views of relevance do not inherently negate the objective and topical view of relevance. On the contrary, if we believe that relevance is a multidimensional concept, the various definitions of the concept may be perceived as reflecting different aspects of relevance. On the other hand, if we extend the dynamic view of relevance, we may also infer that at different stages of the relevance judgment process, different

dimensions of relevance determine document selection, and each stage hence produces distinctive judgment patterns. To apply Swanson's concept of frames of references, we may argue that Frame of Reference 1 does not contradict Frame of Reference 2, and that people are likely to apply Frame of Reference 2 (topical relatedness) in conjunction with Frame of Reference 1 (cognitive newness) in the process of document evaluation. A tentative process model of relevance judgments is proposed in the next chapter. The Process Model sees relevance evaluation as composed of two stages: evaluation of bibliographic records (Stage 1) and evaluation of full-text documents (Stage 2). It is hypothesized that during the first stage users mainly employ Frame of Reference 2, or topical relevance, but as they move to the second stage they apply Frame of Reference 1, or cognitive newness, together with other situational considerations.

Taking various points of view together, it is clear that theories of relevance have evolved to a level where relevance may be defined as a multidimensional dynamic construct that embodies not only objective topical aboutness but also cognitive subjective aspects and situational factors. This conceptual realization is profoundly reinforced through the study of users' relevance criteria which is reviewed in the following section and which indicates that users employ multiple criteria to decide the relevance of a document.

Users' Relevance Criteria

One of the main themes in the empirical literature of relevance is the study of users' criteria of relevance. This line of research typically investigates the criteria that users apply to decide the relevance of a document surrogate or a full-text document. Schamber (1991) 's dissertation research initiated this stream of research and following her example, numerous studies elicited users' criteria in real-life information seeking processes. Table 2.2 outlines the findings of some major empirical works. (For studies that group criteria into categories, only the categorical levels of criteria are listed in the table below.)

Table 2.2
Studies of Users' Relevance Criteria

Authors	Criteria Categories
Schamber (1991)	Accuracy Currency Specificity Geographic proximity Reliability Accessibility Verifiability Clarity Dynamism Presentation Quality
Park (1992)	Internal (experience) context category External (search) context category Problem (content) context category
Cool , Belkin, & Kanter (1993)	Topic Content/Information Format Presentation Values Oneself
Barry (1994)	Criteria pertaining to the information content of documents Criteria pertaining to the user's previous experience and background Criteria pertaining to the user's beliefs and preferences Criteria pertaining to other information and sources within the information environment Criteria pertaining to the sources of documents Criteria pertaining to the document as a physical entity Criteria pertaining to the user's situation
Wang (1994)	Topicality Orientation/Level Subject Area Novelty Expected Quality Recency Reading Time Availability Special Requisite Authority Relation/Origin
Schamber & Bateman (1996)	Aboutness Currency Availability Clarity Credibility

Table 2.2
Studies of Users' Relevance Criteria (Cont.)

Authors	Criteria Categories	
Tang & Solomon (1998)	Topical Relatedness Types of Articles Similar Topical Focus Duplicates Recency Length Depth/Breadth Geographical Focus Version of Article	
Bateman (1998a, 1998b)	Topicality Availability Novelty Currency Quality of Information Presentation Characteristics Source Characteristics Information Characteristics	Factor Analysis Results: Information Quality Information Credibility Information Completeness Information Topicality Information Availability Information Currency

Linda Schamber's (1991) study used open-ended time-line interviews with 30 users from three weather information fields. The users were asked to describe one recent job-related situation in which they required information about the weather in order to make a decision to perform a task. Respondents named the sources they consulted for answering the question they had in their situations. As a result of the content analysis, Schamber identified 22 criteria, and grouped them into ten summary-level categories.

In Schamber's study, most of the categories contain sublevel criteria. Only the first category, *Accuracy*, and the fourth category, *Geographic Proximity*, do not contain subelements. *Accuracy* is defined as "information is accurate," whereas *Geographic Proximity* means that "information covers a certain geographic area." The criterion *Currency* includes a subcategory "Time Frame." *Currency* indicates

that “information is up-to-date or timely.” The next category *Specificity* suggests that “information is specific to user’s need,” and it is represented by two subcategories, “Summary/Interpretation,” and “Variety/Volume.” *Reliability* is reflected by “Expertise,” “Directly Observed,” “Source Confidence,” and “Consistency,” and it is related to the fact that the respondent trusts the source, has confidence in the source and that the source is reputable. *Accessibility*, meaning “source is both available and easy to use,” holds three elements, “availability,” “usability,” and “affordability.” *Verifiability* is instanced by “Source Agreement,” whereas *Clarity* includes “Verbal Clarity” and “Visual Clarity.” *Clarity* in general suggests that “information is presented clearly; little effort to read or understand.” The next category *Dynamism*, refers to “presentation of information is dynamic, active, or live,” and it is projected by the variables of “Interactivity,” “Tracking/Projection,” and “Zooming.” Finally, *Presentation Quality* is defined as “source presents information in a certain format or style, or offers output in a way that is helpful, desirable, or preferable,” and it consists of subcategories “Human quality,” “Nonweather Information,” “Permanence,” “Presentation Preference,” “Entertainment Value,” and “Choice of Format.” It should be noted that Schamber’s study did not involve actual document evaluation.

In her dissertation research, Park (1992) interviewed ten respondents while they evaluated the relevance of bibliographic records in relation to their information problems. As a result, she generated a macro model of relevance, which grouped users’ criteria into three broad categories: *internal*, *external*, and *problem contexts*.

According to Park, the *internal context* category “indicates that a user’s interpretation and decision on a citation seems to stem from his or her own experience or perceptions in the information problem area. The *internal context* is considered to be prior events or beliefs a user has in his or her mind that are not necessarily connected to this search for information” (p. 89). Specifically, the *internal context* includes five elements: User’s previous experience and perceptions; user’s level of expertise in the problem area; user’s previous research experience; and user’s education (or training). The *external context* contrasts with the internal context in that “the origination of the context stems from an individual’s perceptions and situations in relation to the current search, research, and information sources at hand” (p. 89). Park points out that there are six criteria included in the category of *external context*: perception about the search quality; purpose of search (or search goal); perception about the availability of information; priority of information needs; stage of research; and finally, end product of the research. The last category, *problem context*, is the “content-oriented context” and indicates that “the various implied uses of information in relation to expanding one’s ideas and constructing his or her knowledge in the problem area” (p. 90). This category includes the following variables: same (similar) problem area, for definitions; same (similar) problem area, as background; same (similar) problem area, for the methodology; similar problem area, off the target; different problem area, for the methodology; different problem area, for the framework; different problem area, as an analogy; different problem area, as background; different problem area, not of interest; new

information, in the problem context; old (e.g., repetitive) information, in the problem context; and insufficient information, in the problem context. Apparently, the *problem context* is closely associated with topicality or information content, whereas the *internal context* relates to the users' cognition and *external context* reveals the situational aspect of the judgments.

In 1993, Cool, Belkin, and Kanter conducted two studies investigating the factors that influence people's judgments of the relevance of full-text documents. They combined the results of a quantitative study and a qualitative study and offered six facets of relevance. The first is *Topic*, which is described as "how a document relates to a person's interest." Examples of this criterion include "defines the topic itself," "on/not on the topic," "focus (directly on topic, or not)," "part of topic," "treatment," and "important." The second is *Content/Information*, which is differentiated from the *Topic* facet in that it is the "characterization of what it is 'in' the document itself." Examples of *Content/Information* include "basic concepts," "facts/factual," "explanation," "examples," "definitions," "connections," "description," "reasons," "ideas," "tips," "guidelines," "technical knowledge," "interview," "(About) people," "variety," "point of view," "survey," "history," and "level of detail." The third category *Format* indicates the "formal characteristics of the document." Format includes elements such as "lists," "diagrams," "statistics," "pictures," "class text," "book review," "title," "introduction," "division into topics." The fourth facet *Presentation* has to do with how a document is written or presented. Comments related to this category consist of "organization," "matter-of-

factness,” “precision,” “writing style,” “understandability,” “technicality,” “scientificness,” “simplicity/complexity.” The category *Values* consists of dimensions of judgments that are “modifiers of other facets.” Authors suggest that the *Values* category is comprised of “interest,” “amount (lot/little),” “specificity (specific/general),” “goodness,” “usefulness,” “age (of document),” “entertainment value,” “precision (precise/vague),” “bias,” “authority.” The last facet *Oneself* is the “relationship between person’s situation and the other facets.” The elements of such a facet are “need,” “utility,” “desire (‘want’),” “like,” “teaches,” “informs,” “supports understanding,” and “use to which document will be put.” It is interesting that the authors group *Topical* and *Content/Information* as separate categories.

Carol Barry (1993) conducted an empirical investigation eliciting users’ criteria as they read citations and full-text documents. She categorizes 23 criteria into seven general classes. The first class *Criteria Pertaining to Information Content of Documents* includes criteria such as “Depth/scope,” “Objective Accuracy,” “Clarity,” “Recency,” “Tangibility,” and “Effectiveness.” The second class *Criteria Pertaining to Sources of Documents* is specified by “Source Quality” and “Source Reputation/Visibility.” The third category *Criteria Pertaining to the Document as a Physical Entity* contains criteria of “Obtainability” and “Cost.” The fourth category *Criteria Pertaining to Other Information or Sources within the Environment* holds four variables: “Consensus within the Field,” “External Verification,” “Availability within the Environment,” and “Personal Availability.”

There are two criteria under the category *Criteria Pertaining to the User's Situation*-- "Time Constraints" and "Relationship with Author." The *Criteria Pertaining to the User's Beliefs and Preferences* include criteria of "Subjective Accuracy/Validity" and "Affectiveness." Finally, the last category *Criteria Pertaining to the User's Previous Experience or Background* is specified by the following five criteria:

- Background/Experience: the degree of knowledge with which the user approaches information, as indicated by mentions of background or experience
- Ability to Understand: the user's judgment that he/she will be able to understand or follow the information presented
- Content Novelty: the extent to which the information presented is novel to the user
- Source Novelty: the extent to which a source of the document is novel to the user
- Document Novelty: the extent to which the document itself is novel to the user

Note that Barry's last category is semantically equivalent to Park's "Internal Context" category. Her first category *Criteria Pertaining to Information Content of Documents* is similar to Park's *Problem Context* category, except Park's variables are more oriented towards the topicality of the document, whereas Barry's category is leaning towards the quality of the information as perceived by the participants.

The emergence of more and more studies of users' criteria presents a need for reaching a consensus on the grouping and labeling of elicited criteria. As a result, Barry and Schamber produced a paper which compares Barry's findings with Schamber's study and composed a composite set of criteria based on the criteria that are common to both studies. Barry and Schamber (1998) concluded that 11 criteria are shared between two studies: *Depth/Scope/Specificity*, *Accuracy/Validity*, *Clarity*, *Currency*, *Tangibility*, *Quality of Sources*, *Accessibility*, *Availability of Information/Sources of Information*, *Verification*, *Affectiveness*. While Barry and Schamber's paper intends to serve as a concatenation of criterion items, it seems that there remains a need to obtain a synthesized list not merely on the micro individual criterion level but on the macro criteria class/category level where criteria are clustered into meaningful dimensions. Also missing is a sense of *why*, *where*, and *when* criteria are employed in the document evaluation process.

Based on previous research on relevance, Peiling Wang (1994) proposes a cognitive model of document selection. Using that model, she investigated the cognitive aspect of end-users' document selections and the processes of decision making. She found that 11 criteria were employed by the users in selecting document surrogates.

The top criterion in Wang's list is *Topicality*, and it is defined as "what the document is talking about and what the user sees the topic to be with respect to what he/she needs for the task at hand" (p. 46). *Orientation/Level* suggests "at which level the document is, and to what kind of audience it is intended" (p. 47). *Subject*

area is defined as “whether the document falls into the broader area which the user is working/interested in, or whether the author is working in the area the user is interested in” (p. 47). *Novelty* refers to “whether the document or its content is new to the user or whether it has been seen before and is a known item, or its content or information is known to the user” (p. 126). *Expected Quality* is the “perceived or expected quality of a document as judged by personal experience” (p. 48). *Recency* relates to “how long ago was the document published”(p. 49), and *Reading Time* depends on “the user estimates whether he or she has time to read the document, not how long it will take to read it” (p. 49). *Availability* indicates “whether the publication is available from personal collection, from a colleague, from the local library, via interlibrary loan, by ordering from the publisher, or not available at all” (p. 49), and *Special Requisite* means “whether or not some additional equipment or skills are needed to use the document” (p. 50). *Authority* denotes “whether or not the document is written by someone who is recognized in the field,” and *Relation/Origin* suggests that “the origin of the document has a special impact on user with regard to his/her situation because of some pre-existing relationship” (p. 50).

Wang presented two major results with regard to relevance criteria. First, in terms of frequency of mention, *Topicality*, *Orientation*, and *Quality* are the three most frequently used criteria in evaluating bibliographic surrogates. As the most important and basic criteria, *Topicality* accounts for 65% of the total mentioning. *Orientation* and *quality* are both accounted for 9%. Wang states that criteria such as

Novelty, *Subject Area*, and *Recency* are also very critical in relevance evaluations. Second, in terms of the relationship between the DIE (Document Information Elements) available in the bibliographic surrogates and the criteria, Wang (1994) found that topicality was “mainly judged from title, abstract, geographical location, and occasionally from descriptors, journal, and author” (p. 181). She also found that “Orientation was mainly inferred from title, abstract, and journal; and possibly from author. . . Quality was mostly connected to author and journal . . . Novelty was identified basically by title and author” (p. 181). Wang’s study has a profound impact on studies of relevance because the study presents a concrete and operational conceptual model that permits the investigation of cognitive aspects of relevance evaluation during the process of document selection.

In 1997, White and Wang reported a study as a follow-up of Wang’s dissertation research and they found that there were six new criteria when the actual documents are read. *Cognitive Requisite*, which is defined as “whether the user has the knowledge to understand a document;” *Actual quality*, which refers to “the actual quality” as “judged by reading the whole document” (White & Wang, 1997, p. 18-19); *Classic/founder* suggests that “the document is recognized in the field as the first substantial work on a topic or technique;” *Well-known/standard reference*, meaning “the concepts in the document are well-known to the field; or the document is used as text book” (p. 19); *Prolific author* is “when an author wrote many documents on a topic, users may take this situation into account when

reading or citing;” and the *Judge*, which is associated with the “sense of the person who will read or approve the finished product” (p. 19).

The two new citing criteria found by White and Wang are: *Norm*, as in the “perceived expectation or practice in the field for the finished product,” and *Credential*, which “usually refers to paper authored by user and is included primarily to support his own appearance of expertise” (p. 19).

It is worth pointing out that the new reading and citing criteria were found in addition to the 11 original selecting criteria. To be specific, the criteria set for reading consists of not only the six new criteria but also ten of the selecting criteria, excluding only one original criterion, *Special Requisite*. In the citing stage, the criteria that were used include *Topicality*, *Novelty*, *Recency*, *Orientation*, *Authority*, and *Relationship* from the Selecting Criteria, and *Actual Quality*, *Classic/Finder*, *Well-known/Standard*, *Prolific Author*, *Judge* from the Reading Criteria. The Citing Criteria do not contain criterion items such as *Subject Area*, *Expected Quality*, *Availability*, *Time*, *Special Requisite*, *Cognitive Requisite*, and *Reference*. In other words, the criterion that is unique to the stage of selecting is *Special Requisite*; and *Cognitive Requisite* is specific to the stage of reading. For the stage of citing, the unique criteria are the two new criteria: *Norm* and *Credential*.

In a recent study, Schamber and Bateman (1996) conducted a series of validation tests on a collection of 100 criteria accumulated from previous research and in the end they generated five major criteria groups with a total of 23 criteria. The first group is *Aboutness*, within this group the criteria include “about my topic,”

“appropriate,” “pertinent,” “relevant,” and “usable.” The second group is *Currency*, which contains elements of “current,” “recent,” and “up-to-date.” The third group is *Availability*, which is made up by “available,” “accessible,” “convenient,” and “easy to get.” The fourth group *Clarity* consists of factors such as “clear,” “readable,” and “understandable.” The final group *Credibility* is represented by “credible,” “expert,” “I know the publication,” “I know the source,” “prominent,” “reliable,” “reputable,” and “well-written.” This study reveals a strong research motivation to reorganize the criterion list into a meaningful structure. However, it is a little disappointing that the resulting grouping does not seem to be conceptually concrete nor empirically useful as a research protocol for studies to come. For example, the first group “aboutness” contains elements such as “about my topic,” “relevant,” “usable,” and “pertinent.” We might argue that it is possible that when users see a document as “usable,” they not necessarily mean that the document is topically relevant. It is therefore questionable whether such a categorization can serve as a operational coding scheme for other research.

In a study of one person’s relevance judgments, Tang and Solomon (1998) found that at the stage of record evaluation, the subject used a variety of criteria such as *Topical Relatedness*, *Types of Articles*, *Recency*, *Length* and *Language*. The most frequently applied criterion was *Topical Relatedness*, which accounted for 68% of the total mentioning; the second most applied criterion was *Types of Articles*, which accounted for about 14% of the total mentioning. Such a study especially if applied

to more subjects begins to map how relevance criteria are employed as people move through a judgment process.

In her dissertation research, Bateman (1998a) regrouped Schamber and Bateman's (1996) criterion list into nine broad categories. The first category *Topicality* is specified by "about my topic;" the second category *Availability* is represented by the factors of "easy to obtain," and "free or inexpensive." *Novelty* is indicated by "unique or the only source," "original," "new to me," and "familiar." *Currency* is reflected solely by documents being "current," whereas *Quality of Information* include multiple components such as "well-written," "creditable," "accurate," "understandable," "consistent," and "focused." The sixth category *Presentation Characteristics* consists of "presentation of information," "suitable length," "comprehensive," "suitably general or specific," "detailed," "introductory," and "overview." The category *Source Characteristics* refers to the factors such as "I know the author personally," "I know the source," "reputable," "format of the source," and "interactive." The category *Information Characteristics* is defined by "describes methods/techniques," "provides examples," "provides graphics," "statistical approach," "research approach," "provides proof," "controversial," "provides bibliography or links," and "provides background or history." The last category *Appeal of Information* is reflected by comments such as "I like it," "validates my viewpoint," "interesting," and "enjoyable."

Bateman found no significant change in the importance of criteria across the six stages of information searching and the stages of searching, obtaining, and

reading information. She suggests that further study needed to be performed to focus on the use of criteria for partially relevant items or irrelevant items. Another interesting result that Bateman reported is users' ranking of the importance of the criteria in a list of 40. The result of factor analysis indicates that "topicality," "current," "understandable," and "accurate" were rated as the first four important criteria. A mail survey resulted in a slightly different order in the ranking—"accurate," "focused," "topicality," and "understandable" are listed as from the most important to the fourth important. Two additional findings are worth mentioning here. First, "topicality" was rated both in the survey and in the factor analysis as one of the four most important criteria for relevance. This resonates with the results from several previous studies. For example, Barry found that *Information Content* is the most frequently mention criterion category. Wang (1994) also found that *Topicality* was most frequently used criterion, accounting for 65% of the total frequency. *Topical Relatedness* was mentioned 54 times by the subject in Tang and Solomon (1998)'s study, which alone accounted for 68% of the total mentioning. Second, "understandable" was also voted twice as one of the top four important criterion. It could imply that users' cognitive requisite weighs heavily in the use of criteria.

Bateman (1998b) also conducted a factor analysis for the 11 variables that had high importance ratings from the survey data. This led to a three-construct model of high relevance. In her terms, construct *Information Quality* was loaded on by the criteria "Current," "Well-written," "Understandable," "Consistent," and "Focused."

Construct *Information Credibility* included criteria “About my topic,” “Credible,” and “Accurate.” The last construct *Information Completeness* was loaded with criteria “Comprehensive,” “Suitably general or specific,” and “Detailed.”

In examining the factor loading values, Bateman (1998b) argues that “About my topic” seems to separate itself into an independent construct *Information Topicality (Aboutness)*, although Bateman indicates that “the components of topicality or aboutness are often very situational and may be difficult to measure across users” (p. 86). The criterion “Current” was found to be weakly correlated with the construct *Information Quality*, and therefore Bateman proposed that it “probably is a separate dimension of high relevance” (p. 87). The author also suggests that the two additional criteria “Easy to obtain” and “Free and inexpensive” make up another construct *Information Availability*.

To add all the factors together, Bateman in fact presented a model that consists of six constructs or dimensions: *Information Quality, Information Credibility, Information Completeness, Information Topicality, Information Currency, and Information Availability*.

The research on relevance criteria provide strong empirical evidence indicating that users apply a variety of criteria during their relevance decision making. There are three issues resulting from an assessment of the current state of research that are worth noting. First, except for White and Wang’s study, the criteria studies have not distinguished the criteria employed for evaluating bibliographic records from the ones applied for evaluating full-text documents. The

second issue is that there is no consensus as to the categorization and labeling of the classes of criteria. As a result, under different frameworks, a criterion would be named differently and perhaps grouped under different categories. For instance, the criterion item “understandable” is named in Wang’s framework as the criterion “cognitive requisite,” it is grouped with Barry’s *Criteria Pertaining to the User’s Previous Experience or Background* as the “ability to understand.” In Schamber’s structure, “understandable” is a feature of *Clarity*, whereas Cool, Belkin & Kanter (1993) define “understandability” as a result of the *Quality of Presentation*; they indicate that it also has to do with the category of *Oneself*. Finally, in Bateman (1998a, 1998b)’s work, “Understandable” is considered as an element of *Information Quality* in her a priori classification. This research intends to inspect the specific roles of topicality, cognition, and format/quality of information. The conceptual framework presented in Chapter 3 is synthesized from criteria categories that are pertinent to these three dimensions.

The third issue is the methodological aspect of the research. While this issue will be discussed in Chapter 4, it is useful to note that with almost no exception, the empirical studies of the criteria that people employ are elicited by questioning them. In other words, relevance criteria have been elicited from users, mostly retrospectively, drawing from their own judgment experience. Very few studies imposed either an experimental control structure or observed actual behavior. Cool et al.’s (1993) study was the only case that incorporated an experimental design with an associated qualitative study. While the methods that have been used have

resulted in a very comprehensive collection of possible criteria that users apply to make relevance judgments, we are at the point now where we can begin to test these grounded theories with experimental methods that promote greater internal validity.

Relevance Judgments and Formats of Documents

The third body of literature that is related to the dissertation topic consists of the empirical investigations of change in users' relevance judgments across different formats of documents. There are a good number of studies comparing relevance ratings for different types of document representations and documents. Table 2.3 lists several major empirical works, and each study is described by the authors, types of judgments, forms of documents used and main results.

Table 2.3
Change in Relevance Judgments on Formats of Documents

Authors	Types of Judgments	Forms of Documents Used	Major Results
Rath, G. J., Resnick, A. & Savage, T. R.(1961)	Decisions were made based on whether the documents would aid, might aid, would not aid in answering the questions listed in questionnaire	<ul style="list-style-type: none"> • Title • Automatic Abstract • Pseudo-Auto-Abstract • Text 	<ul style="list-style-type: none"> • The percentage of the document scanned: Title: 100%; Auto-abstract: 85% Pseudo-abstract: 78%; Text: 64% • Confidence in usefulness of material: Pseudo-abstract: 90%; Auto-abstract: 86% Text: 86%; Title: 76% • Performance: (perfect acceptance rate 23%) Auto-abstract: 25%; Text: 30% Pseudo-abstract: 16%; Title: 38% • The use of titles in document searching without any additional abstract seems to lead to a higher number of acceptance rate • There is no major difference between the text and abstract groups in their ability of picking the appropriate documents
Resnick A. & Savage, T. R. (1964)	Relevance judgments	<ul style="list-style-type: none"> • Documents-- Documents • Citations -- Citations • Abstracts -- Abstracts • Index Terms -- Index Terms 	Consistency was found in users' judgments of relevance and this consistency seemed to be independent on the kind of materials upon which this judgment is based, except in the case of abstracts
Kent, A, et al. (1967)	Relevance judgments	<ul style="list-style-type: none"> ▪ Citation ▪ Abstract ▪ First Paragraph ▪ Last Paragraph ▪ First and Last Paragraphs 	<ul style="list-style-type: none"> ▪ The relevance decisions made by quasi-motivated users are statistically significantly different from those made by motivated users ▪ With motivated users, extracts, particularly the first and last paragraph combination, serve as well as or even better than conventional output of citation and abstracts

Table 2.3
Change in Relevance Judgments on Formats of Documents (Cont.)

Authors	Types of Judgments	Forms of Documents Used	Major Results
Rees & Schultz (1967)	Relevance judgments	<ul style="list-style-type: none"> ▪ Titles ▪ Citations ▪ Full-text 	There is a general trend for relevance rating decrease from titles to citations to full-texts, although not all the experimental groups shared the same trend. One third of the documents showed some degree of increase in relevance and usefulness ratings across the three representations.
Saracevic, T. (1969)	Three-categorical judgments: Relevant, Partially Relevant, Not Relevant	<ul style="list-style-type: none"> • Title • Abstract • Full-text 	<ul style="list-style-type: none"> • Different representations of documents significantly affect the users' relevance judgment • Immutability: <ul style="list-style-type: none"> Title --> Full-text 85% Abstract --> Full-text 90% Title --> Abstract --> Full-text 78%
Thompson, C. (1973)	Document disposition time; Relevance judgment	<ul style="list-style-type: none"> • Documents without abstracts • Documents with abstracts 	The presence or absence of the abstract made no difference in disposition time, no difference in relevance decision making
Marcus, R.S. Kugel, P., & Benenfeld, A. R. (1978)	Usefulness: (highly useful, somewhat useful, not useful)	<ul style="list-style-type: none"> ▪ Title ▪ Matching Subjects ▪ Subjects ▪ Abstract 	<p>Usability: 1. Subjects, 2. Abstract, excerpts, title, 3. text, 4. Match</p> <p>Utility: 1. Abstract, 2. Title, 3. Subjects, 4. Table of Content</p> <p>Indicativity: Abstract .730; Subjects .704 Matching Subjects .672; Title .637</p> <p>Length Hypothesis: The indicativity of a field of information is positively correlated with its length</p>

Table 2.3
Change in Relevance Judgments on Formats of Documents (Cont.)

Authors	Types of Judgments	Forms of Documents Used	Major Results
Janes, J. W. (1991)	Relevance judgments using magnitude estimation technique	TAB (Title/Abstract/Bibliography) TAI (Title/Abstract/Indexing) TBA (Title/Bibliography/Abstract) TIA (Title/Indexing/Abstract)	<ul style="list-style-type: none"> ▪ large swings under abstract: a substantial change in estimation of relevance based on adding the abstract ▪ "small but frequent" movement under abstract several users frequently changed their judgments after seeing abstracts ▪ stability under bibliographic or indexing information: many subjects exhibited little or no movement of judgment when presented with this new information, producing flat judgment lines ▪ Considerable movement under bibliographic or indexing. Without having seen the abstract, a user may find bibliographic or indexing information more important or useful in making relevance judgment ▪ Five effects: binary sets, ceiling effects, floor effects, increasing trend and decreasing trend are relatively uncommon
Barry, C. L. (1998)	Relevance Judgments	<ul style="list-style-type: none"> ▪ Title ▪ Note ▪ Abstract ▪ Indexing Terms ▪ Full-texts 	<ul style="list-style-type: none"> ▪ Abstracts, titles, and bibliographic citations may contain potential clues to more categories of relevance criteria than indexing terms. Indexing terms address only the topicality of the items, whereas title and abstracts goes beyond simple topicality ▪ Bibliographic information co-occurred with more relevance criterion categories ▪ The clues contained within document representations may depend less on the document representation itself than on the user's context, both in terms of user's previous knowledge and the specific qualities being sought by the user.

Table 2.3 arranges this collection of studies chronologically. The first experimental study, conducted by Rath, Resnick, & Savage (1961), compared the differences in relevance judgments based on four types of documents: title, automatic abstract, pseudo-auto-abstract, and full-text. Automatic abstracts were created through selecting a subset of representative sentences from the documents through an algorithm based on word frequency and distribution. Pseudo-auto-abstracts were generated by selecting sentences from the first 5% and the last 5% of the articles. Both the automatic abstracts and pseudo-auto-abstracts hold 10% of the total length of their full-text counterparts. The investigators found that in terms of the proportion of the documents scanned, titles were read completely (100%), auto-abstracts and pseudo-abstracts were scanned above 75%, whereas the full-texts were only scanned 64%. In terms of judges' confidence in the usefulness of the material, the greatest confidence was found for pseudo-abstracts group, both text group and automatic abstract group indicated 86% of confidence, and title group indicated lowest confidence, which is 76%. In terms of accepting and rejecting documents, the title group had the highest acceptance rate of 38%, while text group had 30%, auto-abstract had 25%, and pseudo-abstract group accepted only 16% of the materials. The authors suggest that using title without abstract lead to a high number of Type II errors, that is, "accepting documents which should be rejected, as not enough information is available to judge the pertinence of documents" (p. 129). The authors conclude that texts and abstracts hold apparent advantage over titles. However, no major differences were found between the text and abstract groups in their ability to

pick appropriate documents, although text group did obtain a higher score than any other groups.

In a study examining the consistency of human relevance judgments, Resnick and Savage (1964) compared users' judgments at two points in time on the same items with a period of a month in between. They grouped the participants into four groups, and each group evaluated one form of documents. The document types include (a) citations, (b) abstracts, (c) index terms, and (d) total texts. It was found that except for the abstract group, the groups demonstrated a strong consistent judgment pattern. The authors indicated that there is no simple explanation as to why judgments on abstracts were not as consistent as for the other three forms of documents; they suggest that the plausible answers may come from additional experimentation.

In 1967, Kent Allen and his colleagues conducted an empirical investigation to address two research questions: 1) are quasi-motivated users (people who volunteer to participate by scanning a list of queries that was previously used in the investigation. People who have no relationship with or knowledge of the original motivated users) a different population than the motivated users (people with real needs of information)? 2) for motivated users, are there other types of cues, in addition to the traditional citation and abstract, that may predict the relevance of original source documents as well as conventional cues?

The investigators employed five types of document surrogates, or, according to authors' own terminology, Intermediate Response Product (IRP). Those IRPs

were citation, abstract, first paragraph, last paragraph, and first and last paragraph combination.

With regard to the first question, the authors found that the correlation coefficient is not statistically different from zero correlation at the 0.05 level of confidence. This supports the idea that motivated and quasi-motivated users are indeed two different populations. With regard to the second question, the authors found that as a form of IRP, the three types of extracts (first paragraph, last paragraph, first and last paragraph) worked better than citation and abstract. Among the three, the first and last paragraph combination performed the best. This form of extract was found among other four IRPs: a) predicted relevance best; b) predicted non-relevance best; and c) had the smallest probability of producing mismatch in judgments.

Kent et al. (1967) thus conclude,

...the relevancy decisions made by quasi-motivated users and those made by motivated users are not correlated significantly above chance expectation. With the motivated users, it was found that extracts (particularly the first and last paragraph combination), functioning as cues to the information content of documents, serve as well as (in our opinion, better than) the traditional intermediated output of citation and abstract. (p. 198)

Document representation is one of the independent variables in Rees and Schultz (1967)'s study of relevance. The authors compared participants' judgments from titles to citations to full-texts, and obtained two interesting results. Firstly, they observed that "it is apparent that the use of brief representations of documents, such as titles and citations, yields evaluations of relevance and of usefulness, which

may differ from evaluations based on full-texts. It is also apparent from the present data that additional factors must be considered in evaluating the effect of representation upon relevance ratings" (p. 164). Secondly, Rees and Schultz discovered that given progressively more information, the medical expert finds subtle distinctions that make a document irrelevant, while the intermediary finds more reasons why the document might be relevant. Rees and Schultz found that although there is a common decreasing pattern in the ratings from title to citation to full-text, one third of the documents showed an increase in ratings across the three forms of documents. From the perspective of the judges, scientifically-oriented judges such as Medical Experts, Medical Scientists, and Residents presented a statistically significant decrease in ratings from titles to citations to full-texts, whereas Medical Librarians exhibited an overall tendency of increasing the ratings from titles to citations to full-texts. It is interesting to see that experts/non experts display different judgment patterns during the process, and this phenomenon apparently suggests that judges' knowledge states play an important role in their relevance ratings. One of the interpretations for this decreasing/increasing phenomenon is that experts (or users with real information needs, as examined in Tang & Solomon, 1998) tend to be more discriminating in the second stage of the evaluation process, whereas nonexperts (or secondary judges) have the inclination to include more items during the second stage. The second interpretation is that it is difficult for the scientifically-oriented judges to evaluate relevance based on titles alone so they "optimistically" give it a higher rating. However, when they view the full-text they become pickier when they see that it actually is not as relevant as they

imagined it would be when they looked at the titles. On the other hand, non-experts such as Medical Librarians, while lacking in detailed subject knowledge, concentrate on matching the key terms in the request to the titles, citations, and full-texts. Given that full-texts provide a greater quantity of text thereby increasing the chance of having more key terms, Medical Librarians tend to rate the relevance value higher for full-texts. One extension of this second interpretation is that intermediaries, such as medical librarians, tend to be more lenient in their relevance assessments than subject experts; they do not want to risk dismissing a document that might be relevant to a subject expert, for whom they are doing a search, for example.

Saracevic (1969) compared users' relevance rating on titles, abstracts, full-text documents, and he found different forms of documents significantly influenced users' relevance judgments. In particular, Saracevic found that the immutability of the rating from title to full-text is 85%, from abstract to full-text is 90%, from title to abstract to full-text is 78%. In other words, 15% of judgments based on the titles changed when the full-texts were presented, 10% of the judgments changed from reading abstracts to the actual documents, and 22% of the total judgments changed when people viewed titles first, abstracts second, and full-text documents the last. Saracevic suggests that "in general, the shorter the representation in comparison to full-text, the more changes in judgments can be expected." He also found that abstracts were more sensitive, more specific, and more effective than titles, and therefore "in general, if given a choice, the judgment from abstracts should clearly be preferred over the judgment from titles" (p. 298).

Thompson (1973) studied the function of abstracts in users' initial screening of documents. Participants were assigned to one of the two groups: one group was shown documents without abstracts and the other group read documents with abstracts. The author compared the time that the participants took to read documents and the relevance judgments they made, and he found that the presence or absence of the abstract made neither a statistically significant difference in disposition time nor a difference in relevance judgments. Based on his findings, Thompson (1973) suggested that "the insistence on the incorporation of abstracts in a document is not clearly warranted to the extent that the purpose is to save the reader's time or to increase his ability to make judgments of relevance" (p. 274).

Marcus, Kugel and Benenfeld's (1978) study serves as a very good example of a comparative analysis on relevance judgments. The investigators first had the users rate the usability and utility of 50 different kinds of catalog fields and found that both *title* and *abstract* were weighted highly in both "usability" and "utility" measures. Usability is the percentage of users indicating, on the basis of a field's general description, that the field should be used in evaluating documents; utility is the percentage of field occurrences checked by users as actually being useful in evaluating the corresponding documents. Ninety-five percent of the users indicated that titles and abstracts are useful in evaluating documents, and 100% of the time titles and abstracts were checked by users in the actual process as being useful to evaluating the corresponding documents. Next, the authors conducted experiments investigating four major content-indicating fields. These fields are: title, subject

term, matching subjects term, and abstract. Participants were shown four cataloging fields in a random order, and they were instructed to evaluate the materials based on the fields given in regard to whether the article appears to be highly useful, somewhat useful or not useful. Following that, they were asked to evaluate the full-text of the article in a similar manner. The results were examined by “indicativity” of each field as to “how well the information in the field conveys the contents of the document it represents.” In concrete terms, “the indicativity of a field was measured by the fraction of evaluations made on the basis of the information in that field that was the same as those made on the basis of the full-text of that article” (p. 16-21). The authors found that abstract had the highest indicativity of 0.73, whereas title had the lowest indicativity rate (among title, abstract, matching subjects, subjects) of 0.64. The investigators then proposed a “length hypothesis” which states that “the indicativity of a field of information is positively correlated with its length” (p. 21). With the average difference of the indicativity scores for four fields at 0.13, Marcus, Kugel and Benefeld suggest that one may infer that full-text documents include 13% of more information that is attributable to the change in relevance decision making. In other words, the judgments based on the abstract have a 13% possibility of change. The authors also monitored some users' actual processes of evaluations, in terms of the amount of time spent in different portions of the text. They found an average user spent 76% of the time reviewing the body of the text, 14% of the time viewing illustrations, 6.5% of the time on abstract and 3.5% of the time on bibliography. They also found

that 88% of the text time was devoted to check relevance, where the remaining 12% was used to obtain information from document.

Janes' (1991) work on relevance judgments and incremental presentation of documents establishes a good research protocol for empirical study of change in relevance judgments. In particular, the study offers a very effective instrument for measuring change in relevance rating — "Motion Index." Participants were randomly assigned to one of the four groups: TAB (Title-Abstract-Bibliographic citation), TAI (Title-Abstract-Indexing terms), TBA (Title-Bibliographic citation-Abstract), and TIA (Title-Indexing terms-Abstract). Relevance judgments were made using the magnitude estimation technique. Such a technique displays a 100 millimeter long line, and participants made a mark on the line corresponding to their impression of the degree of relevance of that document to their queries, from None to Total. The difference in the relevance rating from one document representation to another was subtracted, and the overall difference constitutes the score of "Motion Index" (MI). MI appears to be a simple yet useful measure for the linear change in judgment. Janes found that as a form of document surrogate, the abstract seems to stimulate a different judgment pattern than other types of representations such as title, indexing, and citation. He found that not only did frequent small changes occur after seeing abstracts, but that 64% of the changed judgments were substantial changes with the "Motion Index" score exceeding 40 (i. e., the score differences are 40 millimeters out of 100). Upon examining the motion index for all possible combinations of document representations, Janes (1991) concluded that "the abstract is the most important and most used single piece of

information in relevance judging. Titles are important, but less so than abstracts. There is then a considerable drop to bibliographic information, and finally, indexing” (p. 643).

A most recent study that investigated relevance judgments on a variety of formats of documents was reported by Carol Barry in 1998. The purpose of Barry’s (1998) study is to identify the clues contained in various document representation that “allow users to determine the presence or absence of traits and/or qualities that determine the relevance of the documents to the user’s situation” (p. 1293). Four types of document representations were examined: Title, Note, Abstract, Indexing Terms. Participants were shown all of the materials for the documents, in a randomized order. Each of them was also shown three full-text articles. Participants were instructed to mark or circle the portion of the text that prompted the reaction to pursue or reject an item. They were then asked to orally comment on their markings. Participants’ responses were coded in one of the three ways: information content only, reference traits only, mentions of categories of relevance criteria. In terms of the proportion of the text marked, 79% of the total number of abstracts were marked, and 67% of the full-texts were marked. Indexing terms and title were the third and fourth most frequently marked items, respectively.

Barry found that abstracts, titles and full-texts were the only three document formats that co-occurred with all three categories of responses. She, thus, infers that “abstracts and titles have typically performed effectively as document representations, because they offer access to the same three broad categories of information that are also provided by the full-text of documents” (p. 1299). Because

some criteria are inherently tied with source traits, abstracts, titles and full-texts all co-occurred with a total of 14 relevance criterion categories. This indicates that with regard to providing clues to the potential relevance of documents, “the abstracts and titles are, once again, reflecting the same general pattern as that seen for the full-text of document” (p. 1300). From another perspective, Barry suggests there are certain relevance criteria “for which abstracts and titles are a necessary tool for users” (p. 1300). She further argues that the people’s abilities to detect useful relevance clues from document representations are largely dependent on their previous knowledge on the topics sought.

It is interesting that most of the studies reviewed here conclude that the “abstract” is a useful and valuable form of document surrogate. While Saracevic (1969) found significant change in relevance rating from abstracts to full-texts, Barry (1998) discovered that an abstract mimics the full-text in providing the same relevance clues as the full-text. Janes (1991) found that as a type of document representation, the abstract initiates a somewhat different judgment pattern than other forms of representations, whereas Thompson (1973) found that a document with or without an abstract produced no difference in users’ relevance judgments. This leaves open the question of what exactly is an abstract and what elements should go into an abstract.

An abstract is defined by Tibbo (1993) as the condensation of the original, it is neither a paraphrase nor a summarization. It is a “semantic condensation of the original” so that by reading the abstract, readers obtain the “gist” of the full-text and based on that, they decide whether or not it is necessary to further consult the

original. According to the ANSI/ISO standards for abstracts, a typical abstract for scientific literature should include Purpose, Methods, Results, and Conclusions. Tibbo argues that this guideline is inappropriate for literature of other fields, especially humanities, and proposes that abstracting should be discipline-oriented.

Overall, the studies of relevance judgments using different formats of documents provide evidence for change in relevance ratings as people move from one representation to another. For most of the studies that observed change in judgments, there has been no attempt to explore the reasons for such a change. The most popular assumption is that a change in relevance rating is caused by the apparent differences in the quantity (for example, length of the text) and the quality (different textual structures) of the information contained in different forms of the documents. No study has been conducted to examine whether such a difference in judgments is also caused by a progression in thinking, which leads a user to employ different reasoning structures to determine the relevance of a document. The dissertation research intends to explore this progression of thinking possibility among others.

Stages of Relevance Evaluation and Document Selection

The majority of relevance studies do not consider users' judgments as a process that consists of several distinctive sequences of actions. There are, however, a number of important works on relevance and information seeking that either constructed a process model of information searching or incorporated evaluation stages as a variable in the experimental design.

In their landmark work on relevance, Rees and Schultz (1967) consider the impact of the time element in relevance judgments. They believe that judgments made at different points in time in relation to the work performed will be differentiated. As a result, Research Stages (RS) was included as one of five independent variables. A research stage is operationally defined as occupying “a portion of the total time span of a given research project” and encompassing “a number of related functions performed within that project” (Rees & Schultz, 1967, p. 25). For Rees and Schultz, there are three research stages: RS₀ is described as “encompassing that period of time within which the research problem is formulated;” RS₁ is defined as “encompassing that period of time within which the experimental work is performed;” and RS₂ is defined as “encompassing that period of time within which the data are analyzed and interpreted and conclusions reported” (p. 25). Rees and Schultz found that research stages significantly influence ratings of relevance for individual documents and for the documents as a whole. The interaction between the research stages and judgmental groups was found to be non-significant. It should be noted that in Rees and Schultz’s study the research stages are stages of a research process, they are not intended as the stages that describe the process of document selection.

As a result of a series of longitudinal studies covering a variety of situations of information seeking and searching, Kuhlthau (1993) proposed a model of the *Information Search Process* (ISP), which she characterizes as consisting of seven general stages: *Initiation, Selection, Exploration, Formulation, Collection, Closure, and*

Presentation. Each of the seven stages is accompanied by different features along three dimensions: the affective (feelings), the cognitive (thoughts), and the physical (actions). Table 2.4 illustrates in detail the components of Kuhlthau's ISP model.

Table 2.4
Model of the Information Search process

Stages	Task Initiation	Topic Selection	Prefocus Exploration	Focus Formulation	Information Collection	Search Closure	Start Writing
Feeling	uncertainty	optimism	confusion, frustration, and doubt	clarity	sense of direction/ confidence	relief	satisfaction or dissatisfaction
Thoughts		ambiguity	----->specificity				
				----->			
				increased interest			
Actions	seeking relevant information				----->seeking pertinent information		

(Source: Kuhlthau (1993) Figure 3-1, p.43)

The first stage in the ISP is *Task Initiation* where a person starts to realize that there is a need to seek more knowledge, information and understanding in order to solve a problem at hand. The second stage *Topic Selection* is where people “identify and select the general topic to be investigated or the approach to be pursued” (p. 43). As people move into the *Prefocus Exploration* stage, they “investigate information on the general topic in order to extend personal understanding and to form a focus” (p. 46). The *Focus Formulation* stage involves the development of a focus based on the information encountered. After a focus is gained, the next stage is *Information Collection* when people begin to gather information pertaining to the focused topic. This is the stage when people actively interact with retrieval systems by collecting and evaluating useful materials. At the stages of *Search Closure* and

Start Writing, people complete the search and prepare the final product of the search.

Kuhlthau observed that because of the movement from a general topic to a focus within the topic, people are oriented to different levels of relevance judgments in the content of information received. Specifically, Kuhlthau points out that "in the early stages students sought relevant information related to a general topic. After gaining the focus they sought pertinent information related to the focused topic" (p. 39). In her ISP model, the *Information Collection* takes place after *Focus Formulation*. Hence the action of document evaluation is centered on "seeking pertinent information," as opposed to "seeking relevant information," which happens prior to the formulation of a focused topic.

Kuhlthau's ISP model provides a useful framework for studying movement in the process of relevance evaluation. Since it is difficult to pinpoint at what stages surrogates and documents are sought and evaluated, the ISP stages do not map directly to the document selection/evaluation process. It appears that *Information Collection* is the time period when most document evaluations takes place, but it could also be true that during the early search stage some bibliographic representations are evaluated to help the user to gain a topical focus. Most full-text documents are evaluated at the stage when the searcher has roughly developed a focus and has a sense of direction in collecting and differentiating information. It should also be noted that the ISP model is idealized to simplify the searching

process as a linear, nonrecursive progress. In reality, an information search process is often nonlinear and repetitive.

Several studies investigated relevance judgments using stages as a unit of analysis. For example, Smithson (1994) compared relevance evaluations of three stages—at the end of the online search, at the end of the research project, and at the point of dissertation report—and found changes in judgments. Participants made initial judgments while they evaluated the retrieved records, and then they presented final judgments as they reviewed full-text documents. The final dissertation products of these search processes were analyzed by examining the documents cited.

White and Wang (1997) conducted observations of researcher's document selection processes, and compared the criteria their subjects employed at three different stages: *Selecting* (bibliographic surrogates), *Reading* (scanning or reading a document), and *Citing* (citing the document in the written product). The authors found that participants employ additional criteria both at the stage of reading and citing.

In studying the dynamic nature of relevance judgments, Tang and Solomon (1998) viewed one end-user's relevance judgments at two stages: *record evaluation* and *document evaluation*. The document evaluation stage was further divided into two substages: initial evaluation (quickly scan full-text documents at the time of obtaining the items) and final evaluation (read and study the full-text documents in

detail). Changes in relevance perceptions were found within both the stage of record evaluation and the stage of document evaluation.

Bateman (1998a, 1998b) presented an interesting structure for staging users' relevance judgments. The use of relevance criteria was cross-examined for two time frames: Kuhlthau's seven stages of ISP and a three-stage search process with *Searching*, *Obtaining*, and *Reading* stages. Responses to a survey instrument indicated that the activity of "searching" was connected more with the ISP's *Prefocus Exploration*, *Topic Selection*, and *Information Collection*. The action of "obtaining" was linked highly with the ISP's *Information Collection*, *Prefocus Exploration*, and *Focus Formulation*. "Reading" was identified as mapping to the *Search Closure* and *Information Collection* stages. Bateman's analysis also suggests that if "searching" is viewed as the stage when most surrogate evaluation takes place, the individual who is examining surrogates is likely to be oriented towards "seeking relevant information." This is because the user is still at the stage of *Prefocus Exploration* and *Topic Selection*. If the stages of "obtaining" and "reading" are assumed to be the periods when full-text documents are read, then users are now "seeking pertinent information."

Overall, the cross-examination of relationships among the stages of these two frames by Bateman suggests that the ISP stages as defined by Kuhlthau is not suitable as a framework for describing the peculiarities of the relevance judgment process. Yet, there is a need to build a simple, operational stage structure that describes users' relevance judgments in the process of document selection.

It seems that the best way to approach the staging of relevance judgments is to follow both the Smithson and White and Wang's models. With this approach, the process of document selection is viewed as consisting of three stages: the evaluation of bibliographic records, the evaluation of actual documents, and the citing of documents in the writing products. This approach is used here with a focus on the first two stages of the document selection process.

Summary

This chapter presented a review of related literature. The review centers on four themes: theories of relevance, users' relevance criteria, relevance judgments and formats of documents, and stages and processes of relevance judgments. The theories of relevance considered lead to the idea that relevance is a multivariate construct that contains not only an objective and topical dimension, but also a cognitive and subjective aspect as well as situational elements. It is argued that only an integrated, multidimensional conception of relevance can fully reflect people's relevance judgments during the document selection process. Studies of people's criteria have led to a rich and diversified set of criteria that while conceptually insightful, needs construction and synthesis to produce criteria and classes of criteria to advance IR system design aims. Consequently, the research here focuses on the use of criteria on both a micro individual criteria level and a macro criteria classes level.

With regard to whether users' relevance ratings change as a result of differing document representations, the literature presents conflicting results with

some studies revealing significant change in ratings as users move from one form of representation to another. Yet other studies did not find the change to be significant. From another perspective, most of the studies suggest that the abstract is a very useful and productive form of document representation, with an indicativity rate of 0.73 and immutability of 0.90. Abstracts seem to generate a significantly different judgment pattern than titles or any other types of document representations. It will be interesting to see whether users employ the same criteria to evaluate document surrogates (with a majority containing abstracts) as they do to evaluate full-text articles. In terms of stages in relevance evaluation, a document selection process is considered as moving through three distinctive time points: evaluating bibliographic records, evaluating full-text documents, and citing documents in a written product. Such a structure seems to offer the benefits of simplicity and effectiveness in capturing changes in relevance judgments. The dissertation research examines the first two stages of this process.

Chapter 3

CONCEPTUAL FRAMEWORK AND RESEARCH QUESTIONS

Introduction

This chapter outlines the conceptual framework for the research, which includes a tentative process model and a reconstructed macro level categorization of people's criteria for making relevance judgments. The research questions are then presented. The research questions are developed at two levels: a micro level of individual criteria and a macro level of dimensions or classes of criteria.

Conceptual Framework

Process Model of Relevance Judgment

The review of the literature suggests that relevance is a multivariate construct. Relevance carries at least three connotations: it is objective and topical, subjective and cognitive, and dynamic and situational. Different facets of relevance manifest themselves during the stages of the process of document selection. At this point it is helpful to discuss briefly the characteristics and nature of judgment as human behavior.

In the domain of psychology, judgment as a human action is considered to be different from the acts of decision and choice. A *judgment* is different from a *decision* in that the former is closely associated with knowledge or the process of knowing and inference whereas a decision would produce a course of action and hence it has direct impact on a person's current life condition. Webster's (1961) defines *decision* as "the act of forming an opinion or deciding upon a course of action" (p. 585); *judgment*, on the other hand, is portrayed as "the mental or intellectual process of forming an opinion or evaluation by discerning or comparing" (p. 1223). The view that a *judgment* is rooted in a cognitive evaluative process while a *decision* is realized through an actual action in life is reinforced by philosopher Lonergan's work. Lonergan (1970) indicates that "both decision and judgment are concerned with actuality; but judgment focuses on the need to complete one's knowledge of an actuality that already exists; while decision is concerned to confer actuality upon a course of action that otherwise will not exist" (p. 613). The second difference between a judgment and a decision is that *judgment* is process-oriented whereas *decision* is outcome-oriented. In studying a *judgment* it is important to map the actual process and examine what stages people go through to reach a judgment. To investigate a *decision*, however, researchers recommend focusing more on the quality of that decision as measured against the principles of rationality (McClelland & Mumpower, 1980).

As a response mode, a *judgment* is distinguished from a *choice*. Researchers, represented by Billing & Scherer (1988) and Westenbergh & Koele (1990), propose

that a *judgment* typically requires a full evaluation of every alternative, whereas a *choice* requires only one alternative to be selected and the rest rejected. As a result, a *judgment* is more “cognitively demanding,” since it “demands an explicit rating on each alternative,” it is thus “more deliberative and requires more time and effort” (Billings & Scherer, 1988, p. 4). *Choice*, on the other hand, can be made “with incomplete evaluation of alternatives,” and hence it carries less cognitive load. Westenberg and Koele (1990) suggest that *judgment* normally involves an attribute-wise search pattern (i.e., search and compare the attributes of a decision factor), or inter-dimensional search (i.e., search and compare within one dimension of the decision factor) pattern. A *choice* often engages in an option-wise search pattern (i.e., search and compare different decision factors), or intra-dimensional search (i.e., search and compare among multiple dimensions of decision factors) pattern. The third difference between a judgment and a choice, as pointed out by Westenberg and Koele, is that a *choice* is usually a dichotomous decision, whereas a *judgment* normally applies a continuous or multilevel scale.

Relevance judgment inherits the characteristics of a general human judgment. Consequently, relevance judgment should be considered as very much a cognitive action. It is process-oriented, involving evaluations on attributes, and it can only be expressed as a point on a continuum. The study of relevance judgments thus needs to be performed by mapping and capturing the actual movements in the evaluation process. Building on the analysis of the empirical literature on relevance judgments, there is a need to develop a process model for relevance. In such a process model of

relevance, the dimensions of relevance are actualized, the criteria that are employed by people as the major reasoning principles at different points in time of the process are established, and the change in judgment is regulated by a number of distinctive stages.

There is some theoretical basis for such a process model. One valuable resource is Bert Boyce's two-stage view of relevance judgments. Boyce (1982) speculates that a user's relevance judgment carries through two stages. At the initial stage, the controlling element is *topicality*; people judge a document from the aspect of whether the topic of the document is related to their requests. At the second stage, however, people begin to seek something beyond topicality: the evaluation of a document is based on how informative it is to them as individuals. Boyce suggests that there are at least two elements of *informativeness*: a document needs to be both *understandable* and *novel* to the user. The understandability and novelty (i.e., providing new information to the user) both relate to the personal state of knowledge of the user. Boyce (1982) further indicates that "in the context of the retrieval system we can say that relevance is composed, of necessity, of both topicality and informativeness ... there is no reason to believe that the most topical document is the most informative, and therefore, it is not, of necessity, the most relevant. The most relevant document should not only be highly topical but most informative as well. Relevance would appear to involve a two-stage judgment: first of topicality and then of informativeness" (p. 106).

To put it differently, Boyce sees relevance as a two-stage judgment. In Stage 1 users rely strongly on the principle of topicality. In Stage 2, users depend more on their own cognitive needs. Boyce (1982) states that the two characteristics of informativeness, i.e., understandability and novelty, are “functions of the knowledge state of the requestor” (p. 106). Boyce also points out that at Stage 2, “a document’s relevance is dependent upon the state of the requestor and the state of the requestor changes with the receipt of each informative document” (p. 106). Boyce emphasizes the interaction between the documents and the a priori knowledge of the user. The a priori knowledge of the user is defined in White and Wang’s framework as the criterion of *Cognitive Requisite*. Recall that White and Wang found Cognitive Requisite to be the new and unique criterion in the stage of reading actual documents. This particular criterion was not found in the earlier stage when the document surrogates are reviewed and selected.

The second useful resource is Kuhlthau’s ISP model and her findings regarding to the transformation of judgment orientation in an information searching process. Kuhlthau (1993) found that during the early stages of information searching, a user typically searches for *relevant* information, while in the later stages, specifically after a focus is obtained, the user is geared towards seeking *pertinent* information. In Kuhlthau’s definition, relevant information is information that has connections with a topic or fits with the topic. She states “relevance is a determination that information relates to or applies to the matter at hand, and has a connection or fits with the topic under investigation. Relevant information has

some bearing upon the research topic and is considered useful in a search for information. . . Irrelevant information is outside of the boundaries of a topic and is considered not useful in a search for information” (p. 39). Kuhlthau defines pertinence as something that is much narrower than relevance. Kuhlthau suggests “Pertinence is a determination that information has a more decisive and significant relationship to a topic than relevance and is related to personal information need. Pertinent information is to the point and contributes to understanding or the solution of a problem” (p. 39).

Although Kuhlthau did not specifically assign relevance as reflecting the topical and objective dimension and pertinence as denoting the cognitive, subjective, situational and dynamic aspects, from her definitions, the conceptual connection between the two are apparent. Kuhlthau’s theory of relevance-pertinence transformation could imply the fact that in the early stage of document evaluation, topicality serves as the major criterion. As users move to a later stage, they become more concerned with whether the document is responsive to their cognitive and situational needs.

A tentative process model of relevance is thus formed. An ordinary information search process, as envisioned by the proposed study, starts with the relevance evaluation of bibliographic records. In current IR practice, a majority of bibliographic records contains primary bibliographic information elements such as title, author, journal, abstract, and descriptors. Some of the records do not contain an abstract, yet others contain full-texts. The evaluation of a document representation of any sort is operationally defined as the first judgment stage (Stage

1). The second judgment stage (Stage 2) encompasses the period when users read the actual complete document and make subsequent evaluations.

During these two stages, if one withholds the multidimensional cognitive concept of relevance while extending both Boyce's two-stage view of relevance and Kuhlthau's theory on relevance-pertinence transition, one may infer that when people initially look at the document surrogates, especially when they look at the titles and abstracts, they give more weight to the topicality of the documents as reflected by the record. As they read the full-text documents, they are able to go beyond the topicality of the document and start to examine additional characteristics from the document, for the purpose of fulfilling their own cognitive and situational needs. In other words, whereas the primary concern for the relevance evaluation of records is whether the document that the record represents is topically related to the information request, the main concentration of the relevance evaluation of full-text documents would move beyond the principle of topicality and center on such elements as the newness, interestingness, understandability, as well as other situational factors such as usefulness and satisfaction.

The makeup of the process model of relevance is roughly portrayed in Table 3.1 below. This model focuses only on the period of *document evaluation*, and does not consider the stage of citing. That is, *document evaluation* only contains two stages, whereas *document selection* is a three-stage process with citing as the third stage.

Stage 1 of document evaluation is the evaluation of bibliographic records. At this stage, a user would draw more upon the topical and objective dimension of the relevance, and the principal relevance criterion a user employs is topicality and aboutness. *Stage 2* is signified by the evaluation of full-text documents. At this stage, the cognitive and situational dimension of relevance becomes the main frame of reference, and users apply additional criteria that are salient to their cognitive states and personal situations. Criteria such as “Understandability,” “Newness,” “Interestingness,” and “Usefulness” weigh heavily in the second stage of document evaluation.

Table 3.1
A Tentative Process Model of Relevance

Judgment Stages	STAGE 1 -- EVALUATION OF BIBLIOGRAPHIC RECORDS	STAGE 2 -- EVALUATION OF FULL-TEXT DOCUMENTS
Relevance Dimensions	Objective and Topical Relevance	Cognitive and Situational Relevance
Principal Criteria	Topicality and Aboutness	Cognitive State Utility (e.g., Newness, Understandability, Interestingness, and Usefulness)

Macro Level Recategorization of Relevance Criteria

As analyzed in the previous chapter, the empirical literature on users' criteria has cumulated a rich collection of criteria for relevance. However, due to the diversified structuring of criteria, there are an overwhelming number of criterion items. Criteria have been given different names and assigned to different categories by different researchers. The conclusion is that there is both a conceptual and an empirical need to find the commonality not only at a micro level of individual criteria but most importantly, at a macro level categorization of criteria. There is

also a need for synthesizing and condensing the criteria list so that the overall structure emphasizes a few countable dimensions, and each dimension contains attributes that are concrete and operational.

There are two recent papers that identify shared units among important studies of users' criteria. Barry and Schamber (1998) compared their earlier studies and obtained 11 common criteria. These criteria are: *Depth/Scope/Specificity*, *Accuracy/Validity*, *Clarity*, *Currency*, *Tangibility*, *Quality of Sources*, *Accessibility*, *Availability of Information/Sources of Information*, *Verification*, and *Affectiveness*. Wang (1997) compared criteria from four studies, Barry's, Cool et al.'s, Park's, and Schamber's, against the 11 criteria identified in her 1994 study. It is interesting to analyze Wang's perception of the semantic commonality of criteria across studies. For instance, Barry's *Information Content*, and Schamber's *Specificity* and *Geographic Proximity* were grouped into the category of *Topicality*. Cool et al.'s *Content/Information* is viewed as the same as Wang's *Orientation/Level*. Wang's criterion *Quality* encompassed the meanings of Barry's *Accuracy/Validity*, *Source Quality*, and *Tangibility*, Cool et al.'s *Goodness*, *Usefulness*, *Treatment (depth)*, and *Importance*, and Schamber's *Accuracy*, *Consistency*, *Clarity*, *Dynamism*. Barry's *Background/Experience* and Park's *Users' expertise, prior experience, education* is termed by Wang as *Personal Knowledge*, which is a variable not constructed as a criterion in Wang's model. Wang also notes that different types of documents were used. However, she seems to assume that all criteria are comparable, regardless of the types of documents used. One may argue that since different formats of documents

were used, the comparison would be more meaningful if the aspect of formats of documents was included in the discussion.

A macro level recategorization of users' criteria ought to include criteria that are found most important by the user and used most frequently by the user. These criteria would provide valuable insights to the design of IR systems. With little exception, *Topicality* is found to be both important and used frequently. Wang (1994) found it to be the most frequently used criterion and Bateman's participants rated it as one of the most important criteria. Barry (1994) found that *Information Content* was mentioned most frequently. Since under Wang's grouping, Barry's *Information Content* belongs to *Topicality*, one may suggest that this supports the importance of *Topicality*. The subject in Tang and Solomon (1998)'s case study, used *Topical Relatedness* most frequently while she reviewed bibliographic records. Other criteria that are found to be important and frequently used include *Quality*, *Novelty*, and *Understandability* (Wang, 1994; Bateman, 1998).

By building on the assumptions of the tentative process model of relevance, I propose to restructure users' criteria along three major dimensions: *Topicality*, *Cognitive State*, and *Quality of Information*. Note that many criteria are not included in this structure (e.g., "Usefulness," "Authority," or "Accessibility"). These omissions are made not because the other criteria are not important, but because the study intended to focus on a reasonable number of individual criteria, which could be structured into several sets of macro level criteria. Criteria such as "Accessibility" or "Obtainability" are controlled in the operationalization of this

study and hence are intentionally omitted from the structure. Table 3.2 displays the three classes/categories/dimensions of criteria and the attributes for each class under consideration.

Table 3.2
Recategorized Classes of Criteria

Categories	TOPICALITY	QUALITY OF INFORMATION	COGNITIVE REQUISITE
Criteria	1. Covers the topic 2. Defines the topic 3. Provides background information 4. Provides factual information and data	1. Subject matter is important 2. Information is timely and up to date 3. Accuracy and trustworthiness 4. Clarity and well-written 5. In-depth presentation of information 6. Unique Approach	1. Understandability 2. Newness 3. Similar to what I know 4. Adds to my knowledge 5. Information is interesting and enjoyable

Before providing a formal discussion on the recategorized criteria, two things should be stated. First, this macro level recategorization is not intended as a comprehensive list of criteria. On the contrary, the purpose is to take a reductionist approach to cluster users' criteria along three dimensions. Second, the label for each individual criterion is offered as a semantic guidance, since each criterion is not named in a precise and restricted manner.

The category *Topicality* has to do with the topical orientation and aboutness of the information, and it contains four criteria. These criteria are “Covers the topic,” “Defines the topic,” “Provides background information,” and “Provides factual information and data.” The category *Quality of Information* is defined as describing the genuine quality of the information contained in the documents. This category includes several criteria that have previously been considered as having to do with

information content, quality of presentation and value of information. The first criterion “Subject matter is important” is also labeled as *Importance* in the literature. The second criterion “Information is timely and up to date” is normally termed as *Recency* or related to criterion *Publication Date*. Here it refers exclusively to the fact that the information presented is timely and up to date and hence is treated as associated with the *Quality of Information*. The next four criteria “Accuracy and trustworthiness,” “Clarity and well-written,” “In-depth presentation of information,” “Unique approach” all pertain to the quality of information.

The third category *Cognitive State* refers to the elements that are associated with knowledge states and cognitive structures. In particular, “Understandability” and “Newness” are included as reflecting the interaction between information in the documents and the knowledge state of the searcher. Elements such as “Similar to what I know” and “Adds to my knowledge” are also considered as defining the *Cognitive State* of the user. The last criterion in this category is “Information is interesting and enjoyable,” which was grouped either as the facet of *Value* in Cool et al.’s scheme, or as the element of “Aesthetic Value” in Schamber (1994)’s classification, or as the criterion of “Appeal of Information” according to Bateman (1998a, 1998b). This criterion is defined under the category *Cognitive State* based on the rationale that it essentially describes the nature of information as perceived by users from a cognitive or affective point of view.

The next section elucidates the research questions of the dissertation study.

Research Questions

The research questions originate from the realization that relevance is multidimensional and that the use of relevance criteria is not only dependent on the stages of evaluation but is also a function of the formats of documents. Specifically, the questions are about the use of criteria across the two stages of document evaluation. The units of analysis are *stages* of document evaluation, *dimensions* of criteria, and *formats* of documents. The research questions are specified both at a micro-individual and macro-dimensional criteria level.

Units of Analysis

Stages of Document Evaluation. The process of document selection is viewed as consisting of three stages. *Stage 1* encompasses the period when bibliographic records are sought and evaluated. *Stage 2* encompasses the period when actual full-text documents are sought, read and evaluated. *Stage 3* encompasses the period when a written product of some sort is produced and selected documents are used and cited in that end product. The focus of this dissertation is on the period of document evaluation, i.e., *Stage 1* and *Stage 2* of the process.

Dimensions of Criteria. On a macro level, the use of criteria is examined along three dimensions: *Topicality*, *Quality of Information*, and *Cognitive State*. It is hypothesized that at Stage 1 and Stage 2 of document evaluation, there is a difference in the role of each dimension of criteria. It is expected that the role of *Topicality* and *Cognitive State* would change as users progress from Stage 1 to Stage 2.

Formats of Documents. Journal articles are focal documents of the dissertation project. There are two major formats of the documents under study: *bibliographic records* and *full-text articles*. *Bibliographic records* take one of the following two types: 1) bibliographic *citation*, which includes title, author, journal name, subject terms (descriptors), etc., and 2) citation with *abstract*. In the document selection process, *bibliographic records* are used in Stage 1; *full-text articles* are used for Stage 2 and Stage 3.

This study intends to investigate the difference between the criteria used for judging relevance of document surrogates and the ones used for judging relevance of full-text documents. This investigation centers on the information embedded in the two formats of documents. Formats of documents, thus, serve as the third unit of analysis.

Research Questions

The research questions regarding the use of criteria are pursued at two levels: a micro level of individual criteria and a macro level of dimensions or classes or categories of criteria.

Research Question 1: Use of Individual Criteria across Stages of Document Evaluation

On the micro level, the study explores the use of individual criteria through ratings of importance of the criteria and frequencies of the criteria used:

- a) What criteria do users rate as the most important and use most frequently at each of the two stages?

- b) For what criteria are there changes in the ratings of importance and frequencies of use from Stage 1 to 2? For what criteria are there no changes in the ratings of importance and frequency of use from Stage 1 to 2?

In testing the validity of the process model on a micro level, several criteria are examined in terms of their patterns of change from Stage 1 to Stage 2.

Specifically,

- c) Do the ratings of importance and frequency of use of the criteria related to *Topicality* decrease from Stage 1 to 2? Do the ratings of importance and frequencies of use of the criteria related to *Cognitive State* increase from Stage 1 to 2?

Research Question 2: Use of Classes of Criteria across Stages of Document

Evaluation

On a macro level, the second research question explores the use and change patterns of classes of criteria through ratings of importance of the criteria classes and frequencies of the criteria classes used:

- a) What criteria classes do users rate as the most important and use most frequently at each of the two stages?
- b) For what criteria classes are there changes in the ratings of importance and frequency of use moving from Stage 1 to 2? For what criteria are there no changes in the ratings of importance and the frequencies of use moving from Stage 1 to 2?

In testing the validity of the process model on a macro level, several criteria classes are examined in terms of the patterns of change from Stage 1 to 2.

Specifically,

c) Do the rating of importance and frequency of use of criteria class

Topicality decrease from Stage 1 to 2? Do the rating of importance and

frequency of use of the criteria class *Cognitive State* increase from Stage 1

to 2?

Independent Variable and Dependent Variable

The variables involved include the *stages* (of evaluation) and the *criteria* (on individual and categorical levels). Since the global hypothesis is that the use of criteria is a function of the stage, the outcome variable or the *dependent variable* of the study is the criteria, examined both from an individual aspect and from a dimensional aspect. The *independent variable* is the two stages in the document evaluation process.

Summary

In this chapter, a conceptual framework is laid out, which is supported by a tentative process model of relevance and a macro level recategorization of users' criteria along dimensions of *Topicality*, *Quality of Information*, and *Cognitive State*. Two general research questions of the study are put forth; for each question several more specific questions are posed. The variables involved include the *stages* (of evaluation), which serve as the independent variable and the *criteria*, which serve as

the dependent variables. The outcome variable the criteria are examined both on an individual level and on a dimensional/class/categorical level. The following chapter will discuss the methodology and research design of the dissertation study.

Chapter 4

METHODOLOGY

Introduction

This chapter describes the methodological approaches to the research questions. It opens with a description of the methodological issues evident in the current research on users' criteria and relevance judgments. The next section discusses the use of methodological pluralism for the empirical investigation of relevance and establishes a rationale for performing a laboratory experiment and naturalistic study to investigate different aspects of the research questions. The remainder of the chapter consists of separate discussions of the research designs for the two projects.

Methodological Issues in Studies of Relevance Judgments

Stated Relevance versus User Relevance

One of the top concerns in the empirical design of relevance judgments is whether the study involves people with real information needs or judges who make relevance judgments on their behalf. It has been pointed out over and over again that relevance studies should utilize real people with actual needs for information. This is because only the real people's evaluations of document representations and

documents would reveal the true progression of judgments that accompany users' subjective interests, cognitive needs and situational dynamics. As early as 1966, Cleverdon and Keen distinguished two forms of relevance judgments: *Stated Relevance* and *User Relevance*. They define *user relevance* as the relevance determined by people who make relevance judgments based on their own information needs. *Stated relevance* "can be determined . . . by anybody with reasonable knowledge of the subject field" (Cleverdon & Keen, 1966, p. 256). Since then research on relevance judgments has recognized the importance of employing *user relevance* instead of *stated relevance* in empirical design. In the cases where the participants were not real users, there were attempts to make justifications for the design or to apply a simulated form of *user relevance*. Rees and Schultz (1967), for instance, applied a field experiment approach to stimulate judges' interest so that a stated request becomes close to a real life information need. Recognizing that motivation is a major force in the relevance assessment process, the authors specifically point out that in the design of the study, "an attempt was made to present the research in such a manner as to stimulate interest in the subjects, to give them the 'feel' of the project as well as the substance, and to make it easy for them to think of themselves as being 'in the shoes' of the investigator" (p. 22-23).

Laboratory Experiment Versus Naturalistic Inquiry

In terms of research setting, studies of relevance have applied the methods of both the laboratory experiment and naturalistic inquiry. In his book *Foundations of Behavioral Research*, Fred N. Kerlinger (1986) provides an analytical account of

approaches including *laboratory experiments*, *field experiments*, and *field studies*. *Field studies* are very similar to the method of *naturalistic inquiry*, and, therefore, in terms of the purposes, strengths and weaknesses, both approaches are interchangeable.

Table 4.1 summaries Kerlinger's comparative analysis.

Table 4.1
Laboratory Experiments Versus Field Studies (Naturalistic Inquiries)

Points of Comparison	Laboratory Experiment	Field Studies (Naturalistic Inquiries)
Description and Purposes	<p>Research studies in which the variance of all or nearly all of the possible influential independent variables not pertinent to the immediate problem of the investigation is kept at a minimum. This is done by isolating the research in a physical situation apart from the routine of ordinary living and by manipulating one or more independent variables under rigorously specified, operationalized, and controlled conditions.</p> <p>The aim of laboratory experiments is to test hypotheses derived from theory, to study the precise interrelations of variables and their operation, and to control variance under research conditions that are uncontaminated by the operations of extraneous variables.</p>	<p>Nonexperimental scientific inquiries aimed at discovering the relations and interactions among sociological, psychological, and educational variables in real social structures</p>
Strengths	<ul style="list-style-type: none"> ▪ Relatively complete control ▪ Use of random assignment and manipulation of independent variables ▪ Precision in measurements ▪ High internal validity 	<ul style="list-style-type: none"> ▪ Realism, closest to real life ▪ Significance ▪ Strength of variables ▪ Theoretical Orientation ▪ Highly heuristic, rich in discovery potential
Weakness	<ul style="list-style-type: none"> ▪ Laboratory effect and refined statistics ▪ Artificiality of the experimental research situation ▪ Lack external validity 	<ul style="list-style-type: none"> ▪ Nonexperimental characteristics make weak statements of relations ▪ Lack of precision in the measurement of variables due to the complexity in field situation ▪ Practical problems: potential difficulties in feasibility, cost, sampling, and time

(Source: Kerlinger, 1986, p. 365-375)

As outlined by Kerlinger, laboratory experiments and field studies have different purposes, the former is aimed at *testing hypotheses* regarding the precise relationship between several variables, the latter is aimed at *discovering* the relations or interactions in “complex social and psychological processes, influences, and changes in lifelike situations” (p. 370). The advantages of laboratory experiments include control, manipulation of the experimental setting and precision in testing and measuring. A main disadvantage of the approach is the lack of realism or the artificiality of the situation. In contrast, field studies have the advantages of realism and a process orientation. However, the lack of precision in measurement serves as a potential weakness of the field study approach.

Similar to the structure of field studies, *naturalistic inquiry* is one of the qualitative research strategies that aims at "studying real-world situations as they unfold naturally; non-manipulative, unobtrusive, and non-controlling; openness to whatever emerges--lack of predetermined constraints on outcomes" (Patton, 1990, Table 2.1, p. 40). Instead of testing some predetermined propositions or hypotheses, the researcher who applies a naturalistic approach is interested in understanding and describing events in their naturally occurring states. Thus, in a naturalistic study the investigator imposes no prior manipulation or control on the study setting and variables, and there are no constraints on what the outcome of the study will be. The purpose is to *discover* or provide insights to the complex nature of a process through its ongoing and multidimensional character.

In the empirical study of users' relevance judgments, Park (1994) suggests that naturalistic inquiry is "particularly appropriate in the understanding of how end-users make selection decisions in accepting or rejecting information produced by a document retrieval system which involves complex phenomena in a real life situations" (p. 139). The true nature of relevance judgments, Park claims, can only be examined by using real users in the actual judgment setting. She, therefore, proposes that to capture the cognitive, contextual, and dynamic dimensions of relevance, empirical research on relevance needs to switch from a traditional laboratory experimental approach to a new naturalistic paradigm of inquiry.

The pros and cons of laboratory experiment and naturalistic inquiry are associated with the purposes of the study, how much is known about the phenomena under study, and whether the researcher wishes to gain the benefit of control and precision or the benefit of realism and theoretical discovery. Robertson (1981) offers the following insight into the dilemma involved in information retrieval experimentation:

In order to answer a specific question or questions unambiguously, a test must be designed as far as possible to exclude any extraneous variations which may confuse the results – hence the idea of conducting experiments under laboratory conditions, with all variables controlled as far as possible. On the other hand, in order to answer questions that relate directly to real problems in the design of retrieval systems, and to provide answers to which will apply in real situations, a test must be conducted (as nearly as possible) in an operational environment.

The conflict between these two aims is a real and continuing one. As a result, a whole spectrum of testing methods has been developed, ranging from pure laboratory experiments to the study of real systems and users in their operational environment. (p. 12)

Research on relevance judgments is challenged by the same paradox as described by Robertson. Consequently, it is appropriate to apply multiple methods in the design of research.

A high percentage of the studies on the topic of relevance judgments, as a function of the formats of documents, employed an experimental approach. On the other hand, most of the researchers examining users' criteria and stages of relevance judgments took a naturalistic approach. Still, some researchers were not satisfied with losing the benefits of either realism or control; they thereby attempted to devise a research methodology that is in between the two extremes. Rees and Schultz (1967), for instance, proposed the *field experimental approach*. They believe that such an empirical methodology captures the underlying processes of relevance judgments.

Rees and Schultz (1967) explain that they intended to apply a research design that fulfills two basic requirements. First, the individual who is making the judgment needed to bring a background of interest appropriate to the evaluation task. The researchers wanted to create a situation where the judgment is meaningful and realistic to the individual. Second, the investigators believe that the design should contain the characteristics of laboratory experiments with controlled variables so that the results can be quantifiable, replicable and generalizable. To do so, they wanted to set the judges in the same situation, using the same research in the same manner and at the same stage. In order to compromise between these two

seemingly incompatible objectives, Rees and Schultz suggest the field experimental approach.

The objective of this approach is to provide an experimental setting with a sufficient appearance of realism and naturalness to make the experiment acceptable to the subject as a meaningful situation within which to react. At the same time, the control and manipulation of a specified variable is possible. In effect, this approach is a combination of a field study – systematic, objective observation in a natural setting – and a formal laboratory experiment involving full specification of the experimental procedures, control of variables and quantification of results. (p. 21-22)

Thus, the field experimental approach in relevance study is a method that attempts to combine realistic relevance judgments with experimental procedure and quantitative data analysis. Such an approach, however, as indicated by Kerlinger (1986), does not have a sharp distinction from a laboratory experiment. Kerlinger suggests that “the differences are mostly matters of degree” (p. 369). The field experiment approach and the laboratory experiment approach are two methods on a continuum rather than holding separate territories. It is therefore arguable that this “in-between” approach is THE approach for the study of relevance.

Current State of Research on Users’ Criteria

Before deciding which approach is appropriate for investigating the research questions here, it is necessary to discuss briefly the state of research on relevance criteria. As pointed out earlier, most of the studies on users’ criteria applied the naturalistic approach. This line of research has generated a great quantity of criteria. Even though many of these criteria resulted from naturalistic studies, the criteria seem to have lost their grounding as they come from long lists of criteria without their context (i.e., mappings to their situation of use). Yet, no consensus on

criteria has been reached. It is time to both map the actual circumstances of criteria use for relevance judgment and to present grounded statements about the context of use of these criteria as definitions of how these criteria are employed. This foundational work could, then, provide a basis for proposing testable hypotheses regarding the variables involved in the process as well as shed light on my expectations regarding change in criteria use from Stage 1 to Stage 2.

As the dissertation research questions hold a process orientation, the naturalistic inquiry seems preferable. The research questions are also concerned with specific relations between variables or sets of variables, for which the laboratory approach has advantages. Therefore, it seems that both laboratory and naturalistic methods are appropriate to investigate the research questions.

Toward a Methodological Pluralism

The Philosophy of Methodological Pluralism

Scientific research has been operating under two major paradigms of inquiry. One is the positivist paradigm, and the other is the phenomenological or interpretative inquiry. These two approaches are based on fundamentally different philosophical assumptions, and thus are perceived as two competing and conflicting approaches to research. Positivism sees reality as an objective entity containing movements and actions that have underlying structures and regularities. The purpose of the scientific investigation is to reveal these structures and regularities via systematic and controlled means. Kerlinger (1986) defines the positivist inquiry as the “systematic, controlled, empirical, and critical investigation

of natural phenomena guided by theory and hypotheses about the presumed relations among such phenomena” (p. 10). The positivist approach is normally associated with the research design of laboratory experiment and quantitative data analysis. As Patton (1990) described, positivism “uses quantitative and experimental methods to test hypothetical-deductive generalizations” (p. 37).

The phenomenological paradigm, on the other hand, is based on the assumption that reality is essentially defined by social beings as they operate in specific social contexts and evolves under human subjectivity. With this understanding, phenomenological or interpretive inquiry applies “qualitative and naturalistic approaches to inductively and holistically understand human experience in context-specific setting” (Patton, 1990, p. 37).

While many scholars have engaged themselves in debate about the general advantages and disadvantages of one paradigm over the other, Patton (1990) pleads for “a paradigm of choices.” He argues that researchers should free themselves from their methodological prejudices or their “one-sided paradigm allegiance.” Instead, the concentration ought to be on how appropriate an approach is to a specific research situation. Patton believes that a one-sided advocate of a paradigm is pointless, and that the most important issue is not to be restricted to a single paradigm but to select the paradigm appropriate to the investigation of specific situations. He further states,

A paradigm of choices rejects methodological orthodoxy in favor of *methodological appropriateness* as the primary criterion for judging methodological quality...The paradigm of choices recognizes that different methods are appropriate for different situations. Situational

responsiveness means designing a study that is appropriate for a specific inquiry situation. (p. 39)

The philosophy of “methodological pluralism” was developed by Wildemuth (1993) based on Patton’s notion of “a paradigm of choices.” Methodological pluralism encourages use of multiple approaches. Wildemuth explains that “in this view, there is no such thing as the one correct scientific method. Instead, the method to be applied in a particular study should be selected based on the research question being addressed” (p. 451). She uses two empirical studies on users’ online search behavior as examples of the methodological pluralism. In both studies, an interpretative research was incorporated with a positivist design. Wildemuth (1993) concluded that an “interpretative approach can be combined effectively with positivist research, in spite of the fact that the two approaches take very different views of the nature of reality and how one comes to know about or understand reality” (p. 466). Wildemuth further concluded that

...neither positivist nor interpretative research can address every research question...However, a positivist approach is helpful in determining whether the theory is generalizable to situations other than those in which it was developed...an interpretative approach was helpful in understanding how the searchers themselves understood those searching behaviors and why they behaved in the way they did. It is hoped that both positivist and interpretative approaches have a place in information and library science research. (p. 466)

This dissertation research adopts the philosophy of methodological pluralism and promotes that both a positivist experimental method and an interpretative naturalistic method are useful to address different aspects of the dissertation research questions.

The Concept of Triangulation

One concrete development in the theory of methodological pluralism is the notion of “triangulation.” In social science research, the use of multiple approaches to one research question is sometimes called “methodological triangulation.”

Singleton, Straits and Straits (1993) state that it is unfortunate that “social researchers rely altogether too frequently on a single method or measure when a number of approaches could be brought to bear on the research question” (p. 393). They contend that “given the limitations and biases inherent in each of the main approaches – indeed, inherent in all research procedures – the best way to study most research topics is to combine methodological approaches” (p. 391).

According to Singleton et al. (1986), the logic of triangulation originated from the field of navigation, when applied to social science research triangulations pertains to situations in which two or more dissimilar measurement approaches are used.

The key to triangulation is the use of *dissimilar* methods or measures, which do not share the same methodological weaknesses – that is, errors and biases. The observations or “scores” produced by each method will ordinarily contain some error. But as the pattern of error varies, as it should with different methods, and if these methods independently produce or “zero-in” on the same findings, then our confidence in the result increases. (p. 392)

It is my belief that the philosophy of methodological pluralism and the concept of triangulation provide a foundation for investigation of the research questions here. Both laboratory and naturalistic approaches have virtues in studying the questions. However, each of them holds some disadvantages that cannot be ameliorated until both approaches are utilized together and the results are

treated as complements of each other. The differences and similarities in the two results should provide insights to the questions.

In the study of relevance judgments, Cool, Belkin, and Kanter (1993)'s research adopted a design incorporating methods of laboratory experiment and qualitative inquiry. In the next section, I will discuss briefly how methodological pluralism and the concept of triangulation are actualized in Cool et al.'s work.

Cool, et al.'s Study: Methodological Pluralism in Relevance Research

This study has two threads. It was aimed at "investigating the factors which underlie people's judgments of the relevance or usefulness of documents to particular information problems" (p. 77). The first study, GMU (George Mason University) study, involved 300 freshmen taking an introductory computer science course. The second study, the Humanities Scholars study, included 11 in-depth interviews with senior level scholars from the fields of history, English and philosophy. In the GMU study, the participants were given an assignment to write an essay and were required to cite a minimum of five sources in the essay. In addition, the students were asked to provide information about their thought process on each document when they looked at the document. Participants in the Humanities Scholars study, in contrast, were interviewed about their typical information seeking tasks, the uses they made of texts, and the judgment processes they engaged in when they evaluated documents for the intended use.

The investigators consciously contrast the two studies by their settings and research designs:

The two studies clearly investigate quite different situations, with quite different methods. The former is concerned with students, facing a predefined task, and is focused on the specific characteristics that they use to judge documents with respect to that task. It addresses these issues by studying a large number of people, and a few categories of data, looking for some regularities in their behaviors, without necessarily addressing the reasons for those behaviors. The latter is concerned with experienced scholars, in their entire scholarly life and associated problems and goals and is focused on their interactions with texts in general. It addresses these issues by studying in depth a relatively small number of people, in their activities in general, and attempts to understand the processes which influence their uses of documents. (p. 79)

The authors further claim that both methods are valid in their own right, and that a complete understanding of relevance judgments can only be accomplished by combining the findings from both studies. The authors also indicate that the reason that they reported the two studies together is that they believed that the two results complement each other.

Cool et al.'s study demonstrates that using an empirical design of multiple methods to investigate users' relevance judgments is not only operationally feasible but also effective in examining complicated issues such as the use of criteria in the process of document evaluation. Based on this understanding, I included both a laboratory experiment component and a naturalistic observation component to the investigation of the research questions.

Design of the Dissertation Research

The dissertation research consists of two parts: a laboratory experiment project and a naturalistic observational case study.

The Laboratory Experiment

Description. The purpose of the project was to test the hypothesis regarding change in the use of criteria as participants moved from Stage 1 (evaluation of bibliographic records) to Stage 2 (evaluation of full-text documents). Participants were instructed to read a set of preselected materials related to the “Year 2000 Problem” (Y2K). The materials were retrieved by conducting searches on the *Computer Select* CD-ROM database. As a result of their reading, participants needed to produce an outline of presentation points for a talk on the Y2K problem and its social effects. They read materials first in the surrogate (abstract) format then in the full-text format. As they read the materials, they selected the items that they believed would be useful for their presentation outlines. They also responded to questionnaires regarding how important each of the 15 reasons (e.g., “The documents present information on the Year 2000 Problem that is up to date,” “The documents contain information that is new to me, or present ideas that I have never come across before”) was in contributing to their decisions. This element of the research obtained approval from UNC Academic Affairs Institutional Review Board (AA-IRB) in February 1999 (Appendix A).

Procedure. Specifically, an experimental session involved three major steps:

Step 1. Participants were informed at the beginning that their task was to write a presentation outline on “The Year 2000 Problem and Its Social Effects” after they read materials on the topic. They then filled out a *Preevaluation Information Sheet* (Appendix C) which collected basic personal information about

the participants such as the class level and major. They also indicated how much they knew about the topic.

Participants then read 20 records containing bibliographic information, e.g., title, author and abstract. Participants selected a number of records that they wanted to read in full, and they indicated their selection decision by checking corresponding article titles on the *Abstract Checklist* (Appendix D). They were then given a questionnaire (Appendix E) containing a list of 15 possible reasons (criteria) for their selections, and they indicated on the questionnaire how important each reason (criterion) was for their decision. The importance of each reason (criterion) is expressed on a scale from 1 (not at all important) to 7 (extremely important). Participants were asked to note that “It is crucial that you understand that you are rating how important each issue was in your decision process, and NOT how true each statement is about the particular summaries that you read.”

Step 2. Participants were provided with the full-text articles corresponding to the items that they selected in Step 1. They read through the articles they chose, and selected the articles that they believed were useful for their writing. They indicated their selection by checking the articles from the *Full-text Checklist*. The *Full-text Checklist* was the same as the *Abstract Checklist*, except that for this step the titles of the articles that participants selected were highlighted and they selected only from the highlighted titles. Next, the

participants completed the *Full-text Questionnaire*, which was the same as the *Abstract Questionnaire*, but this time the added note read

As before, you are rating how important each issue was in your decision process, and NOT how true each statement is about the particular article you read. Also, you are engaged in a new task with a new purpose; you should not feel that your new responses need to be similar to your earlier responses about article summaries.

Step 3. Participants produced a presentation outline (Appendix F is a copy of the presentation outline completed by a participant). During each experimental session, one participant was selected at random to present the outline to the experimental group. The presenting individual received ten dollars. Each experimental session ended with a debriefing about the purpose of the research as the participants were expected to learn the underlying motivation of every experiment they participated in as a member of the Psychology Human Participant Pool.

Participants. The participants were recruited from Human Participant Pool in the Department of Psychology at UNC-CH. The Human Participant Pool contains undergraduate students who are enrolled in Psychology 10 -- Introduction to Psychology. Each Student from that class is required to complete a five hours of experimental credits. The majority of the students in the pool are Freshman or Sophomore.

Pretest and subsequent changes in the design. Three volunteers were invited to participate in the pretest of the experiment, which was held in February 1999.

Two changes were made as a result of that pretest. First, the original number of abstracts to be read was 25. This was reduced to 20 since participants from the pretest indicated that 25 was too many to read within the time frame of two hours. Second, originally the *Abstract Questionnaire* and the *Full-text Questionnaire* were exactly the same. Participants from the pretest commented that this led to the assumption that the task was to recall the first-time ratings. In support of these comments, the pretest data show a very similar ratings on both the abstract and full-text questionnaires. In order to eliminate the possibility of this misconception, the *Full-text Questionnaire* included a separate note to remind participants that they are now engaged in a different type of task.

The Naturalistic Study

Description. The purpose of the naturalistic component of the research was to study the criteria that people employ to evaluate the relevance of documents at two different stages of a research process. By naturalistically observing and recording participants' relevance judgment behaviors in the process of conducting online searches and evaluating documents for the writing of term papers, this project intended to understand the patterns of use of judgment criteria across Stages 1 (evaluation of bibliographic records) and 2 (evaluation of actual full-text documents). The project obtained approval from UNC Academic Affairs Institutional Review Board (AA-IRB) in January 1999 (Appendix B).

Participants. Students enrolled in a graduate-level course in the Department of Psychology at the University of North Carolina at Chapel Hill were invited to

participate in the research. The course was Psychology 284 “Research Synthesis.” It was a course in quantitative meta-analysis. Most of the students in the class are Ph.D. students in the Department of Psychology. In order to obtain a high-pass for the course, students were encouraged to complete a meta-analysis on a topic they were interested in. One participant (Participant 5) who joined the project later, was not enrolled in the class. He is also a psychology Ph.D. student, but instead of working on a meta-analysis project, he planned to work on revising a manuscript.

Procedure. The naturalistic component was based on individual cases.

Specifically, each case proceeded through the following steps:

- Initial interview discussing the topic and preliminary thoughts on the topic.

The interview was semi-structured and was audiotaped.

- Observation of the literature search of online bibliographic retrieval databases such as PsycInfo, Medline, Social Science Citation Index, Science Citation Index, Lexis/Nexis, Legaltrec, whichever the participants chose. The participants were also requested to talk aloud while making selections of bibliographic records. Participants were asked to articulate the reasons for selection decisions. The whole session was audiotaped.

- Collection of copies of full-text articles based on each participant’s selections.

- Each participant was then provided with *Document Evaluation Sheets*

(Appendix G contains some examples of document evaluation sheets completed by the participants) accompanying copies of the actual articles. Participants read the articles at their own convenience (unobserved). Upon reading each article,

participants completed their evaluation for that article on the *Document Evaluation Sheet*. Appendix H includes samples of the document evaluation sheet completed by the participants.

- After finishing the reading of the articles, each participant was interviewed about their reading processes and the comments made on the evaluation sheets. During the meeting, participants orally reviewed each of the articles they read, with the actual articles and document evaluation forms in front of them as references. The oral evaluation session was audiotaped.
- After the oral review of the articles they read, participants were asked to discuss their experience in using criteria. In the end, they all reported on the status of their products. The post evaluation discussion was semi-structured and was audiotaped.

Methods of Data Analysis

The Laboratory Experiment

The purpose of the laboratory experiment was to investigate people's perceptions of the importance of criteria in contributing to their relevance judgments as they move from the stage of record evaluation to the stage of full-text evaluation. The data analysis was performed on the micro- and macro- levels.

Micro Level Analysis

The micro level analysis was based on the computation of average importance ratings for each of the 15 criteria across all the participants. The change

in the importance ratings was measured both by differences in the importance ratings of the criteria between the two stages and by the differences in the ranked positions of the criteria between the two stages. The ranked position of a given criteria in each of the two stages was developed by arranging the rating values in a descending order.

Macro Level Analysis

The macro level analysis was conducted by categorizing the 15 criteria into three broad classes of criteria. Below is the actual listing of criteria as they appear on the questionnaires. The labels included in the parenthesis were abbreviations used as representations of the criteria, as reported in Chapter 5.

1. The documents discuss the Year 2000 Problem and its social effects. (“Discuss Y2K and its social effects”)
2. The documents present information on the Year 2000 Problem that is up to date. (“Information up to date”)
3. The documents provide rich, well-rounded information about the social effects of the Year 2000 Problem. (“Rich, well-rounded information”)
4. The documents provide factual information on the social effects of the Year 2000 Problem, based on analyses of actual data. (“Clear and well-organized information”)
5. The information presented in these documents is understandable, because it is presented in a way that is not too technical or scientifically complex. (“Understandable, not too technically complex”)
6. The information contained in these documents deepens my understanding of the social effects of the Year 2000 Problem. (“Deepen my understanding”).
7. The documents contain interesting information; I enjoyed reading them. (“Interesting and enjoyable”)
8. The documents provide information on the origin and causes of the Year 2000 Problem. (“Cover Y2K origin and causes”)
9. The documents discuss issues that are real and important in our daily lives. (“Issues are real and important”)
10. The information in these documents is presented clearly, and in a well-organized fashion. (“Clear and well-organized information”)
11. The documents present their information and arguments in a manner that is fresh and unique. (“Fresh and unique approach”)

12. The documents contain information that is new to me, or present ideas that I have never come across before. (“New information and new ideas”)
13. The documents seem to provide a clear definition of the Year 2000 Problem. (“Provide definition of Y2K”)
14. The documents provide information that is consistent with what I already know about the social effects of the Year 2000 Problem. (“Consistent with previous knowledge”)
15. It seems likely that the information provided in these documents is accurate and trustworthy. (“Accuracy and trustworthiness”)

The researcher’s a priori classification of criteria is as follows:

Topicality: 1, 4, 8, 13

Quality of Information: 2, 3, 9, 10, 11, 15

Cognitive Requisite: 5, 6, 7, 12, 14

For the macro level analysis, a multivariate approach was employed. Multivariate Techniques include “an assortment of descriptive and inferential techniques that have been developed to handle situations where sets of variables are involved either as predictors or as measures of performance” (Harris, 1975, p. 5). Since the variables involved in this study are sets of variables with a number of variables within each set, multivariate analysis is thus appropriate. Harris (1975) indicates that multivariate analysis techniques have two areas of strength. On the descriptive side, “they provide rules for combining the variables in an optimal way;” on the inferential side, “they provide a solution to the multiple comparison problem” (p. 5-6). Here, Factor Analysis was employed as a multivariate descriptive technique, and Hotelling T^2 test was used for inference about changes in scale scores.

Hotelling T² Test. Hotelling T² test was used since the predictor variable *stage* is a two-level nominal variable whereas the outcome variables *criteria* include three sets of variables with a minimum of four variables within each set. The *Hotelling's T²* statistic is suitable for examining simultaneous differences between the meanings of two or more sets of variables. The following equation calculates the *Hotelling's T²* statistic:

$$T^2 = n(\bar{d})'S^{-1}(\bar{d})$$

where \bar{d} is the vector of mean differences, S is the unbiased sample covariance matrix $\frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})(d_i - \bar{d})'$ and n is the sample size (Krzanowski, 1988; Rencher, 1998).

Factor Analysis. The method of Factor Analysis was used to generate statistically valid clusters of criteria based on the participants' importance ratings. The following reviews some of the important concepts related to the method of Factor Analysis.

Factor Analysis "refers to a variety of statistical techniques whose common objective is to represent a set of variables in terms of a smaller number of hypothetical variables" (Kim, 1978, p. 9). The fundamental assumption of factor analysis is that there are underlying factors responsible for the covariations among the variables. Typically, a Factor Analysis is performed when a person has a large number of variables and wants to obtain a sense of the general dimensions in these variables. The purpose of Factor Analysis, as described by Cattell (1978), is "to find

a new set of variables, fewer in number than the original variables, which express that which is common among the original variables” (p. 16). Guertin and Bailey (1970) define Factor Analysis as “a formal decision making process to explicate subsets of covarying variables no matter how numerous they are” (p. 1).

There are many reasons that a researcher may undertake Factor Analysis. Comery (1973) provides a set of possible scenarios for using Factor Analysis. According to him, an investigator may “have measurements on a collection of variables and would like to have some idea about what constructs might be used to explain the intercorrelations among these variables,” or the investigator wishes to “test a theory about the number and nature of the factor constructs needed to account for the intercorrelations among the variables he is studying.” A third possibility is that the person wants to “determine the effect on the factor constructs brought about by changes in the variables measured and in the conditions under which the measurements are taken” (p. 4). All these scenarios to some extent describe what I intended to do for the dissertation research questions, and, therefore, Factor Analysis was applied to analyzing the experimental data.

Factor Analysis operates on the correlations among variables, and starts with “a matrix expressing the correlations of each variable with every other variable” (Guertin & Bailey, 1970: 1). Factors are extracted based on the correlation coefficient values. These factors are conceived as latent variables that are underlying causes of the observed causes. Each observed variable has “loadings” on the factors, and these loadings reflect the extent to which the variable is related to the hypothetical

factor. The factor solution is further rotated and more factors are extracted until in the end an optimal factor solution is reached.

Regarding the interpretation of the meaning of the factors produced in a factor solution, Guertin and Bailey (1970) present an important point that would be helpful for understanding the data analysis results of this experiment as discussed in Chapter 5: “proper analysis can give factors which are easily named by reference to the nature of the variables heavily loaded on each. Yet the factor definition remains algebraic and the label is merely a mnemonic convenience” (p. 84). In other words, the factor solutions produce factors that represent data patterns from statistical points of view, they are not necessarily meaningful.

In this study, Factor Analysis produced fruitful results, which, will be discussed in full detail in the next chapter. The factor solutions produced through the Factor Analysis technique were compared with the researcher’s a priori classification model. This comparison allowed further conceptual development and led to a revised classification of the criteria.

The Naturalistic Study

The data analysis of the naturalistic observational study started at the micro individual criteria level. The measurement was the frequencies of use of the criteria, comparisons were made among common criteria between the participants’ frequency ratings of Stage 1 and Stage 2. The analysis concerning each participant’s use of criteria at the two stages was first performed. Next, the frequency rates

across all the participants were generated to obtain an overall sense of the use of criteria by the participant group.

Following the micro level examination, a macro level analysis was conducted. The criteria used by the participants were grouped into several classes, and the frequency rates of the use of the criteria classes were recalculated for each of the participants. The analysis concerning each participant's use of criteria classes at the two stages was first performed. Next, the frequency rates across all the participants were generated to obtain an overall sense of the use of criteria class by the participant group.

Participants' perception of their use of criteria was also examined. The participants' own reflections provided the rich context for the use of criteria. Furthermore, their discussion on the importance of the criteria they used at the two stages and the change in the use of criteria provided many interesting and valuable insights into criteria use. These ideas became the main source of inspiration for the researcher's taxonomy of criteria. The results are reported in Chapter 6.

Summary

This chapter discusses the state of research for the study of relevance criteria and some of the main research approaches or design orientations taken by the researchers of relevance. This researcher considers the philosophy of methodological pluralism and the concept of triangulation and, on their basis, argues that multiple methods are appropriate to the investigation of the dissertation questions. Consequently, the research design included a laboratory experiment and

a naturalistic study. The two projects included participants with different backgrounds and research experiences, and used different measures and methods of analysis. The commonalities and differences in the laboratory results and the naturalistic results provided insights to the change in the use of criteria and thus enhanced understanding of the nature of relevance. Reports of the results of the laboratory experiment is presented in Chapter 5, and of the naturalistic study in Chapter 6.

Chapter 5

RESULTS – THE LABORATORY EXPERIMENT

Introduction

The laboratory experiment was designed to investigate the participants' ratings of the importance of relevance criteria at two focal points in time: after they read document surrogates, and after they read full-text articles. Data analysis was conducted first on the micro level, for which average importance ratings of the 15 criterion items were compared between Stage 1 and 2. Following that, a macro level analysis of the ratings of three criteria categories was employed to present a broader, dimensional view of trends in the data.

This chapter focuses on the results of the laboratory experiment. First, a description of the characteristics of the data is presented. Second, a detailed analysis of the data at both the micro (individual criterion) level and the macro (criteria category) level is reported. The analysis also includes a Hotelling T^2 test and Factor Analysis. While the former examines the statistical significance of the change in the ratings of multiple sets of variables according to the researcher's a priori model, the latter generates statistically valid factor solutions for criteria groupings. The chapter ends with a discussion of the experimental findings.

Characteristics of the Data

Participants

The experimental project was conducted in the spring of 1999. The participants were 90 undergraduate students¹ (44 females and 46 males) drawn from the human participant pool in the Department of Psychology of the University of North Carolina at Chapel Hill. The participant pool includes undergraduates who are required to take part in psychological research as a condition of their enrollment in a lower division introductory course; they are allowed to choose the studies in which they participate. Seven experimental sessions were administered; each session had about 12 people. Below is the basic information about the group of participants involved in this study.

Level of Education. As shown in the table below, over half of the participants were freshmen. Sophomores made up more than a quarter of the total participants, whereas juniors and seniors comprised 13% of the entire group. Three participants did not indicate the education level on their information sheets.

Table 5.1
Participants' Levels of Education

Levels	Number of Participants (<i>N</i> = 90)	Percent of Total
Freshman	52	58%
Sophomore	23	26%
Junior	8	9%
Senior	4	4%
Not Specified	3	3%

¹ The actual number of participants is 91; one participant provided invalid data and was excluded from the final analysis.

Major. Participants held a wide spectrum of majors, ranging from Biology, Business, Psychology, to English, Journalism, Political Science, Nursing, Communication, and Environmental Sciences and Engineering. About 12% of the participants had not decided their majors at the time. Table 5.2 displays the top four majors, most of the other majors include only one or two people.

Table 5.2
Participants' Majors

Major	Number of Participants (<i>N</i> = 90)	Percent of Total
Undecided	11	12%
Biology	6	7%
Business	6	7%
Psychology	5	6%

Knowledge level. On the pre-evaluation information sheets, participants selected from a list of six the statement that described their knowledge level on the topic of "The Year 2000 Problem (Y2K) and its Social Effects." Participants were allowed to check more than one option. Table 5.3 summarizes the participants' reports of their knowledge levels. About 48% of the participants indicated that they had heard about the topic frequently, and about 44% of them had discussed the issue with their classmates or with friends and family. While 32% of the participants had only heard about it once or twice, two percent indicated that they had never heard of it. On the other hand, three percent of the participants were involved in some type of advanced research on the issue, and 10% of the participants expressed a concern about the seriousness of the problem.

Table 5.3
Participants' Knowledge Levels on the Topic of Y2K

Knowledge Levels	Number of Participants	Percent of Total
Heard about it a lot through reading newspapers, watching television, or browsing the internet	44	48%
Discussed it formally in class or informally with family members, friends or colleagues	40	44%
Heard about it once or twice through reading newspapers, watching television, or browsing the internet	32	35%
Have been very concerned about it	10	11%
Conducted research on it and wrote a report about it	3	3%
Never Heard of It	2	2%

Types of Data

The experimental project collected two major kinds of data. The *quantitative data* consists of participants' ratings on two questionnaires (one was completed after reading the abstracts, the other after reading the full-text articles). The data was processed and analyzed, and the results are reported below in detail. The selection decisions made by individuals for the two rounds were also recorded; they are also quantitative in nature. The second type of data is *textual data*. Textual data are made up of all the documents reviewed, including 20 abstracts and 20 full-text articles, and most importantly, participants' written products.

Although the analysis of the textual data is beyond the scope of the dissertation, preliminary analysis suggests some interesting issues for the future. For example, by comparing the textual segments in the documents reviewed by an individual and textual units in the presentation outline produced by the person, links can be made to the threads that the participants built between the original

texts and the recreated texts. By studying such textual connections, it may be possible to map the steps involved in the process of composition, and gain some insight into how relevant units of information contained in the original documents were extracted, processed, reorganized, integrated, and finally recreated in a written format.

Another potential research focus would be to study the characteristics of the documents that an individual reviewed and then explore the association between the characteristics of the texts read at the two stages and the person's importance ratings on the 15 criteria for the corresponding text type (abstract or full-text). This analysis would provide a measure of how differences in the ratings were influenced by what documents a participant read.

A third potential point of data analysis is to look at the correlation between participants' knowledge levels and their importance ratings of the criteria. The purpose is to explore whether people with different knowledge backgrounds of the topic held different preferences for the 15 criteria. It is possible to group the participants according to their knowledge levels (e.g., the high knowledge level group and the low knowledge level group), and test whether the ratings of criteria or criteria classes are statistically different from group to group. The result would enrich the findings of the current analysis.

Additional contextual data is provided by participants' comments during the debriefing towards the end of each experimental session. The participants were asked whether they believed that there were changes in the use of criteria. Two people explicitly pointed out that there probably was no change. Others did not

comment on the issue. A majority of people felt very strongly about the deceptiveness of the abstracts. They commented that several full-text articles were not what they had expected after reading the abstracts. “One of the articles I read turned out to be very disappointing,” said one participant, “it doesn’t focus on what the abstract says that it would.”

The next section reports the analysis of the quantitative data (i.e., the questionnaire data).

Results

After reading 20 abstracts on the topic of the Year 2000 Problem, the participants made selections of the articles that they wanted to read in full. Then, they rated the 15 criteria by how important they thought each of the criteria was in contributing to their selection decisions. They did that for a second time after they read the full-text articles that they selected and decided which were the ones that they were going to use for writing the presentation outline. Statistical analysis of the questionnaire data was performed on two levels: the micro level analysis of the ratings of the individual criterion items, and the macro level analysis of the ratings of the criteria categories.

Micro Level Analysis

Within the Stages

The mean ratings at Stage 1 are listed in descending order in Table 5.4. “Understandable, not too technically complex” was rated the highest with a mean of 5.822. “Accuracy and trustworthiness” is the second highest criterion, with a mean of 5.800. The lowest rated criterion was “Consistent with previous knowledge,” with an average rating of 4.033. The range of the mean ratings is about 1.79 (the difference between the highest and the lowest), which is less than two points on a scale of seven.

In interpreting the standard deviation values of the criteria, one thing to keep in mind is that the data here is non-ratio data, and hence it is somewhat questionable to make parametric estimations. However, as a measure of dispersion, the standard deviations provide an indication of the level of agreement among the participants. The standard deviations for Stage 1 criteria have a range from 0.978 to 1.604. The criteria that have relatively higher consensus in the ratings were “Issues are real and important” (SD = 0.978), “Clear presentation of information” (SD = 1.028), “Accuracy and trustworthiness” (SD = 1.029), and “Understandable, not too technically complex” (SD = 1.034). Criteria that have relatively lower consistencies in the ratings were “Consistent with previous knowledge” (SD = 1.604), “Cover Y2K origin and causes” (SD = 1.487), “New information and new ideas” (SD = 1.471), and “Interesting and enjoyable” (SD = 1.452).

Table 5.4
Mean Ratings of Criteria at Stage 1

Criteria	Mean Importance Rating Stage 1	Standard Deviation Stage 1
Understandable, not too technically complex	5.822	1.034
Accuracy and trustworthiness	5.800	1.029
Discuss Y2K and its social effect	5.767	1.218
Issues are real and important	5.622	0.978
Information Up to date	5.544	1.308
Provide definition of Y2K	5.478	1.201
Clear presentation of information	5.433	1.028
Rich, well-rounded information	5.344	1.113
Deepen my understanding	5.278	1.190
Factual information and actual data	5.244	1.376
Interesting and enjoyable	4.778	1.452
Cover Y2K origin and causes	4.633	1.487
New information and new ideas	4.478	1.471
Fresh and unique approach	4.211	1.204
Consistent with previous knowledge	4.033	1.604

As a counterpart of the information presented in Table 5.4, Table 5.5 displays the mean importance rating for Stage 2 after the participants read the full-text articles that they chose at Stage 1. At Stage 2, “Discuss Y2K and its social effect” is the most important criterion. It was rated 6.022, which is a bit higher than the rating at the previous stage. “Issues are real and important” became the second highest, whereas “Accuracy and trustworthiness” was the third highest rated criterion. “Consistent with previous knowledge” remained the lowest at Stage 2, with a mean rating of 4.133, increasing by 0.10 from Stage 1. The difference between the highest and lowest ratings for Stage 2 is 1.89, which is a slightly larger than that for the abstract evaluation.

The standard deviations for Stage 2 range from a low of 0.942 to a high of 1.677. The criteria that have relatively higher agreements among the participants include “Issues are real and important” (SD = 0.942), “Deepen my understanding” (SD = 1.023), “Accuracy and trustworthiness” (SD = 1.034), and “Clear and well-organized information” (SD = 1.083). Criteria that have relatively high discrepancies in the ratings were “Consistent with previous knowledge” (SD = 1.677), “New information and new ideas” (SD = 1.612), and “Cover Y2K origin and causes” (SD = 1.564). Note that for both stages, “Issues are real and important” and “Accuracy and trustworthiness” were among the criteria with relatively high consistencies, whereas “Consistent with previous knowledge,” “New information and new ideas,” and “Cover Y2K origin and causes” were among the criteria with relatively low consistencies in the ratings.

Table 5.5
Mean Ratings of Criteria at Stage 2

Criteria	Mean Importance Rating	Standard Deviation
	Stage 2	Stage 2
Discuss Y2K and its social effect	6.022	1.218
Issues are real and important	6.011	0.942
Accuracy and trustworthiness	5.822	1.034
Understandable, not too technically complex	5.689	1.196
Deepen my understanding	5.622	1.023
Provide definition of Y2K	5.544	1.163
Rich, well-rounded information	5.489	1.220
Factual information and actual data	5.478	1.309
Clear and well-organized information	5.478	1.083
Information up to date	5.400	1.421
Cover Y2K origin and causes	4.956	1.564
Interesting and enjoyable	4.911	1.387
New information and new ideas	4.611	1.612
Fresh and unique approach	4.222	1.436
Consistent with previous knowledge	4.133	1.677

Figure 5.1 provides a comparative overview of the ratings of the criteria across the two stages. As shown in Figure 5.1, the ratings for the two stages followed very similar patterns, especially with regard to the relatively lower rated criteria. The lower rated criteria for both stages include “Consistent with previous knowledge” and “Fresh and unique approach.”

There was a greater degree of change for the highly rated criteria. Nevertheless, four criteria “Discuss Y2K and its social effects,” “Accuracy and trustworthiness,” “Understandable, not too technically complex,” and “Issues are

real and important,” which were all rated the top four most important criteria at Stage 1, were also rated as the top four criteria at Stage 2.

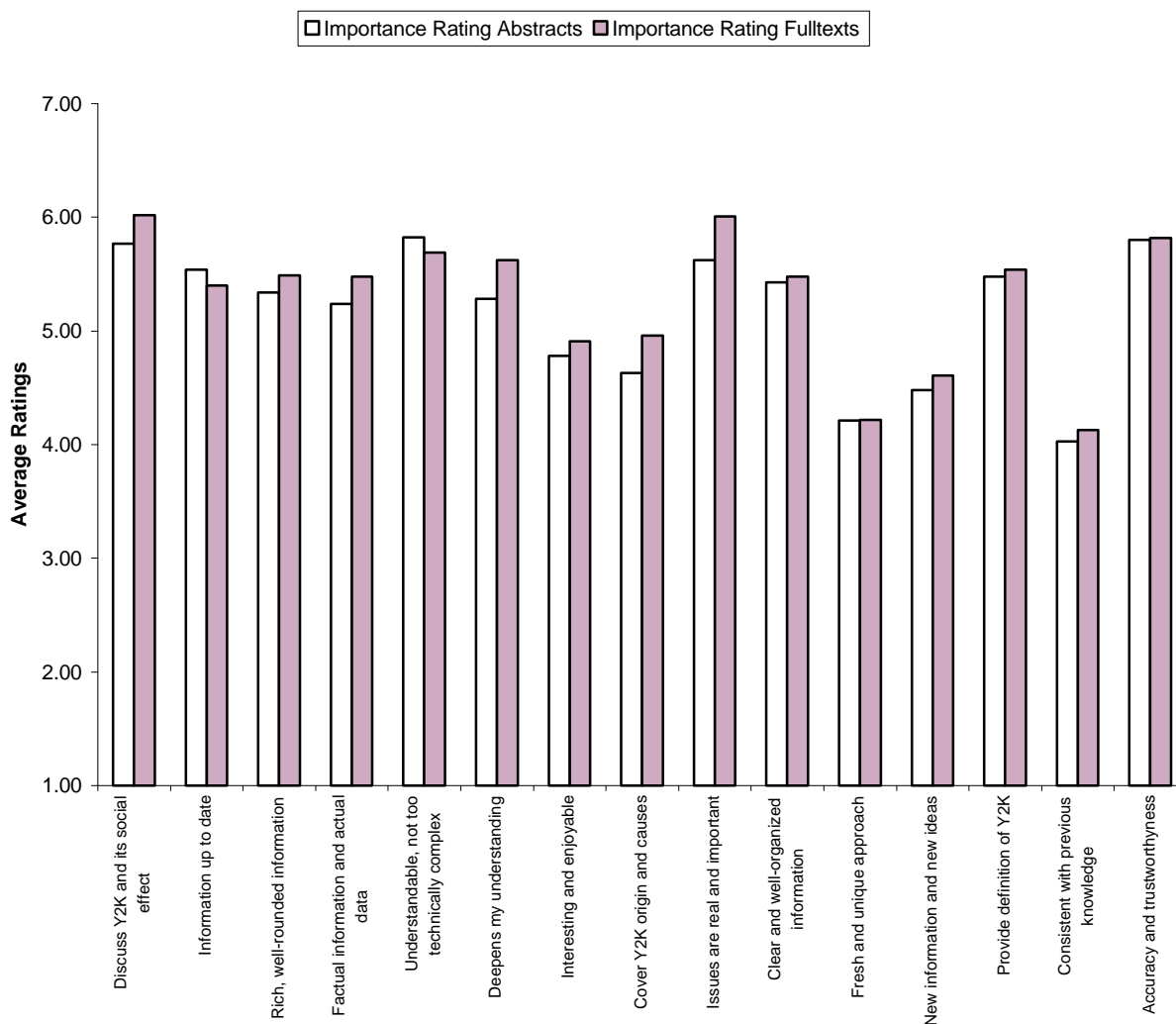


Figure 5.1. Mean Ratings of Criterion Items Stage 1 and Stage 2

Between the Stages

The differences in the mean ratings between the two stages were examined from two perspectives. Firstly, a direct comparison was made between the mean

ratings of the two stages. Secondly, a comparison of the rankings of the criteria was made between the stages.

Figure 5.2 illustrates the difference in mean ratings for Stage 1 and Stage 2. Criteria that are on the left side of the vertical scale are those that were rated higher at Stage 1 than Stage 2. The ones that are on the right side are the items that were rated higher at Stage 2 than Stage 1.

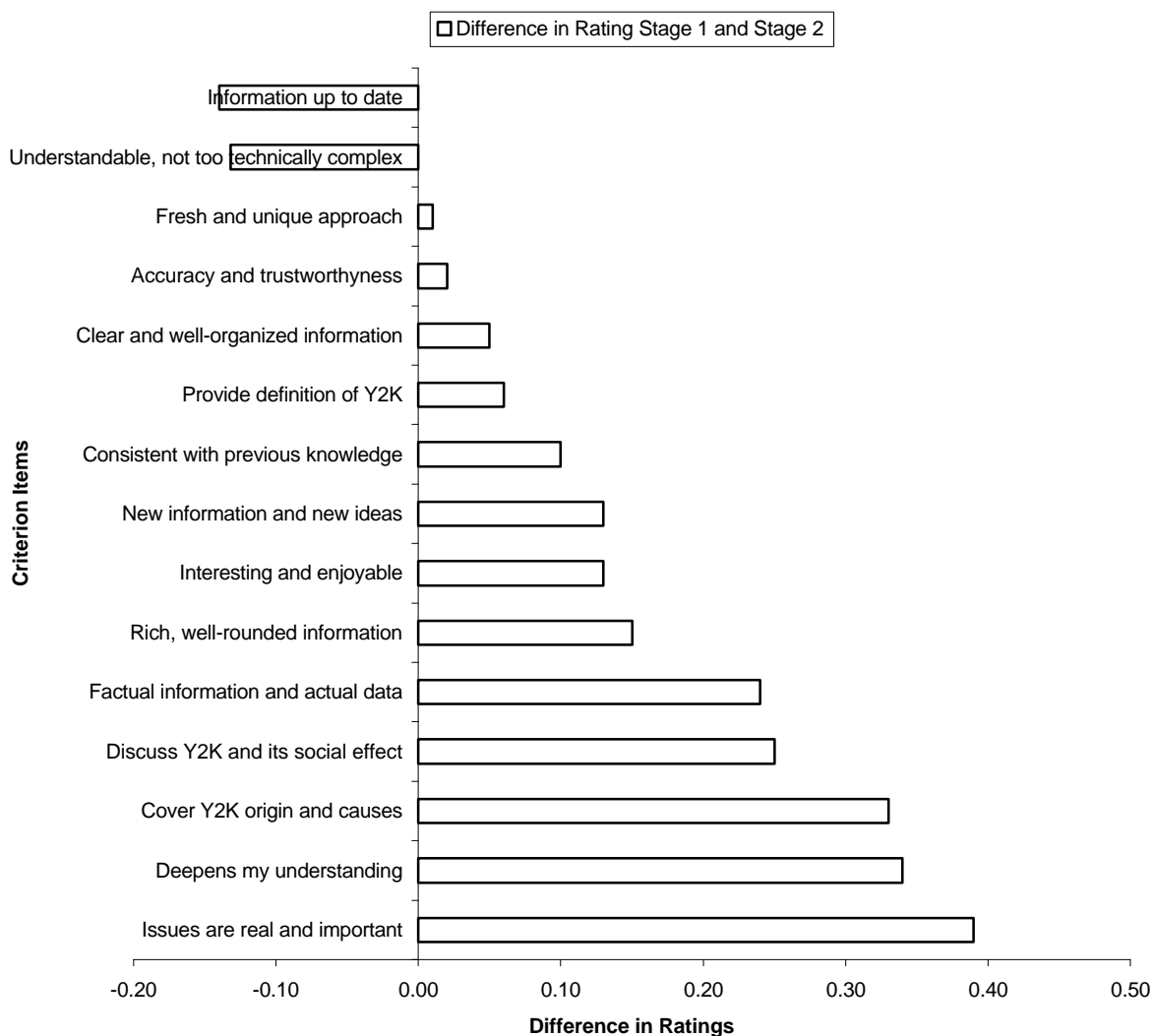


Figure 5.2. Difference in Mean Ratings of Criteria

The differences between the mean ratings of both stages are small. Figure 5.2 shows that the criteria that have the greatest differences (d = mean rating of full-text criterion - mean rating of abstract criterion) from Stage 1 to Stage 2 include “Issues are real and important” ($d = 0.39$), “Deepen my understanding” ($d = 0.34$), “Cover Y2K origin and causes” ($d = 0.33$), “Discuss Y2K and its social effect” ($d = 0.25$), and “Factual information and actual data” ($d = 0.24$). Criteria “Understandable, not too technically complex” and “Information up to date” are the only two whose ratings decreased from Stage 1 to 2. Criteria that have relatively small changes in ratings include “Fresh and unique approach” ($d = 0.01$) and “Accuracy and trustworthiness” ($d = 0.02$). Overall, the differences between the two stages are not large, with the greatest change around 0.40, which is less than a half point on a scale of seven. Nonetheless, the differences do seem to suggest a shift towards criteria that depend on the more in depth explanation available in a full-text over an abstract.

The second perspective that reflects the differences in criteria ratings is a comparison of change in ranking positions for the 15 criteria between the two stages. Table 5.6 displays the rankings. The rankings were generated according to the values of mean criteria ratings within each stage. For instance, “Information up to date” was ranked as number 5 at Stage 1, whereas at Stage 2 it dropped to number 10. “Deepen my understanding” was ranked number 9 at Stage 1, whereas at Stage 2 it advanced to be the fifth in rank. Notice that there are four criteria that had the exact same ranking positions across two stages. These four criteria are

“Provide definition of Y2K,” “New information and new ideas,” “Fresh and unique approach,” and “Consistent with previous knowledge.” The last three criteria are also the three lowest ranked criteria for both stages. The top four criteria for both stages are the same set, although the ranking position varied from Stage 1 to Stage 2.

Table 5.6
Rankings of Criterion Items Stage 1 and Stage 2

Criteria	Rankings Stage1	Rankings Stage2
Understandable, not too technically complex	1	4
Accuracy and trustworthiness	2	3
Discuss Y2K and its social effect	3	1
Issues are real and important	4	2
Information up to date	5	10
Provide definition of Y2K	6	6
Clear and well-organized information	7	9
Rich, well-rounded information	8	7
Deepen my understanding	9	5
Factual information and actual data	10	8
Interesting and enjoyable	11	12
Y2K origin and causes	12	11
New information and new ideas	13	13
Fresh and unique approach	14	14
Consistent with previous knowledge	15	15

For a more graphic illustration, Figure 5.3 draws lines connecting the rankings of the same criteria across two stages. In this figure the top rank is valued as 15, whereas the lowest is scaled as 1.

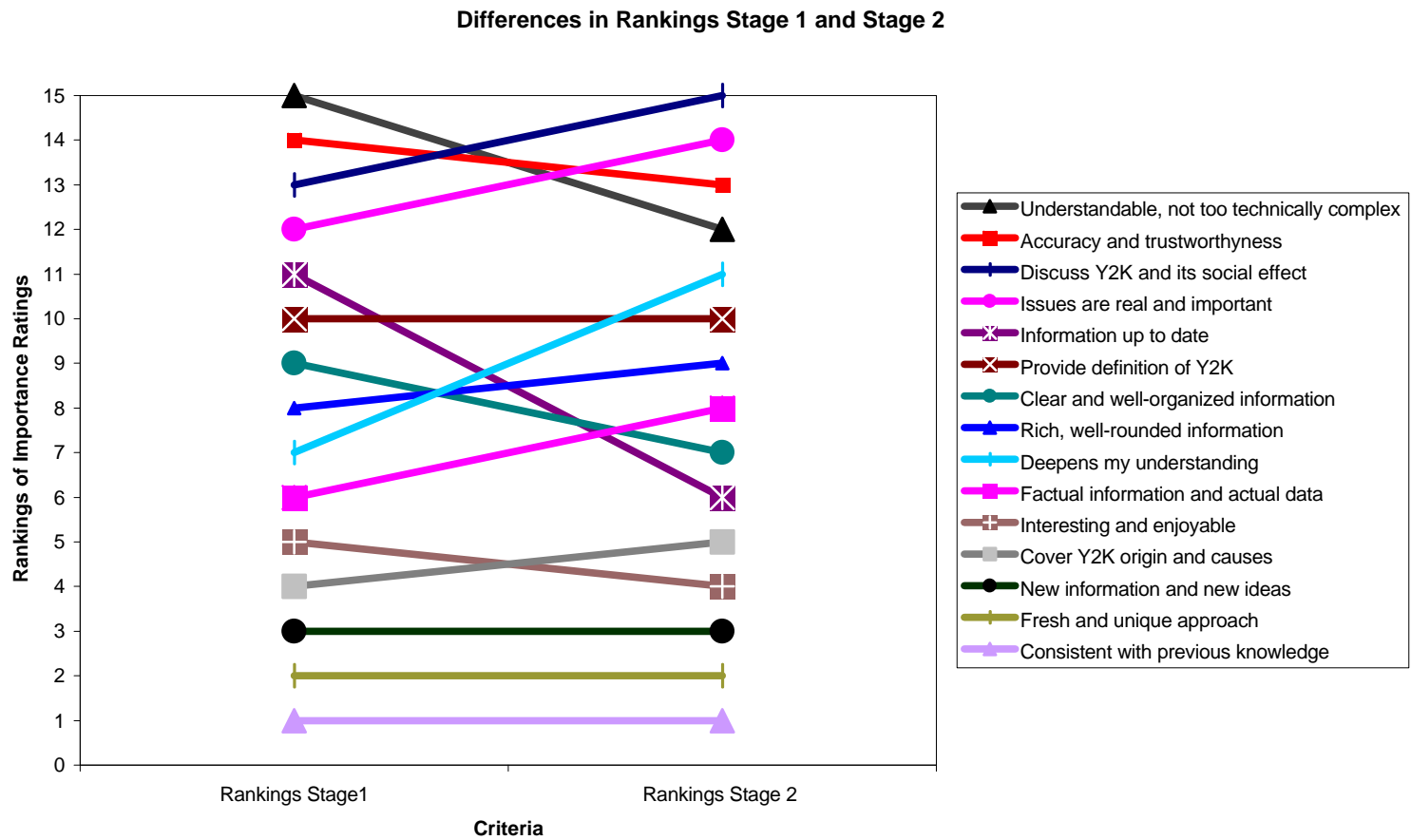


Figure 5.3. Rankings of Mean Criteria Ratings Stage 1 and Stage 2

As is indicated in Table 5.6, the changes in the ratings of the criteria, as measured by their ranked positions, are also relatively small across the two stages. There were four criteria that remained at the same ranked positions between the stages, with d_r equal to zero ($d_r = \text{rank of Stage 2} - \text{rank of Stage 1}$). The greatest change in rank is for “Information up to date” ($d_r = -5$), with its ranked position dropping 5 units at Stage 2. On the other hand, “Deepen my understanding” has an increased ranking ($d_r = 4$) at Stage 2. Another criterion with a relatively large change in ranked position is “Understandable, not too technically complex” ($d_r = -3$). This particular criterion is ranked first in Stage 1 and decreased by 3 units in Stage 2.

Other Criteria

Four participants suggested “other” criteria on their Stage 1 questionnaire (questionnaire completed after reading abstracts). One participant indicated, “I tended to look at the ones that made the effects of the Y2K problem look the worst; basically the ones that make it look like the ‘biggest deal.’” Another participant seemed to have the same feeling; he proposed that an important criterion should be that the document contains “predictions and examples of how severe the social effects may be.” The third participant stated that a desirable document would be the one that “provides an off-the-wall explanation of the effects of the Year 2000 problem,” or “explains how the problem will affect the ‘Ordinary Joe.’” The last comment was that a document that “doesn't focus on software/hardware issues” would be considered relevant. Notice that the first three comments are similar to

the criterion “Issues are real and important,” whereas the fourth comment is to some degree related to “Understandable, not too technically complex.”

After reading the full-text articles, one participant wrote that it is important that an article “saw both sides of the argument on whether or not Y2K will be a big problem.” Another participant believed that the length of articles is important; a good article should be “short and to the point.”

To summarize, several participants agreed that a possible criterion is to find articles that describe the seriousness of the Year 2000 problem. Considering the timing and the nature of the topic, it seems reasonable that the participants would pay extra attention to the impact of the problem. On the other hand, this could indicate that a majority of the publications selected for this project are somewhat inadequate in addressing the seriousness of Y2K. All this may explain why the criterion “Issues are real and important” was rated highly at both stages, even more so for Stage 2 when the full-texts were examined than for Stage 1 when the abstracts were read.

Macro Level Analysis

Researcher’s a priori Classification Model

As outlined in Chapter 3, the purpose of the study is not only to examine the use of relevance criteria at an individual level, but also to consider it on a broad, categorical level. It was proposed that the 15 criteria might be grouped into three classes of criteria: *Topicality*, *Quality of Information* and *Cognitive State*. Table 5.7 below reiterates this classification.

Table 5.7
 Researcher's a priori classification of Criteria

<i>Categories</i>	<i>Group Number</i>	<i>Criterion Items</i>
Topicality	1	Discuss Y2K and its social effect
	1	Provide Definition of Y2K
	1	Factual information and actual data
	1	Cover Y2K origin and causes
Quality of Information	2	Accuracy and trustworthiness
	2	Issues are real and important
	2	Information up to date
	2	Clear and well-organized information
	2	Rich, well-rounded information
	2	Fresh and unique approach
Cognitive State	3	Understandable, not too technically complex
	3	Deepen my understanding
	3	Interesting and enjoyable
	3	New information and new ideas
	3	Consistent with previous knowledge

Importance Ratings for Criteria Categories

The macro level analysis was conducted based on the grouping of the three criteria classes listed in Table 5.7. The mean importance rating for a given class was obtained by averaging the mean importance ratings of the elements involved in the class. The mean importance rating for each class at Stage 1 and 2 and the standard deviations are given in Table 5.8. Figure 5.4 provides a visualization of the mean rating statistics.

Table 5.8
Mean Ratings of Criteria Classes Stage 1 and Stage 2

Criteria Category	Mean Importance Rating Stage 1	Standard Deviation Stage 1	Mean Importance Rating Stage 2	Standard Deviation Stage 2
Topicality	5.281	0.881	5.500	0.882
Quality of Information	5.326	0.686	5.404	0.710
Cognitive State	4.878	0.773	4.993	0.886

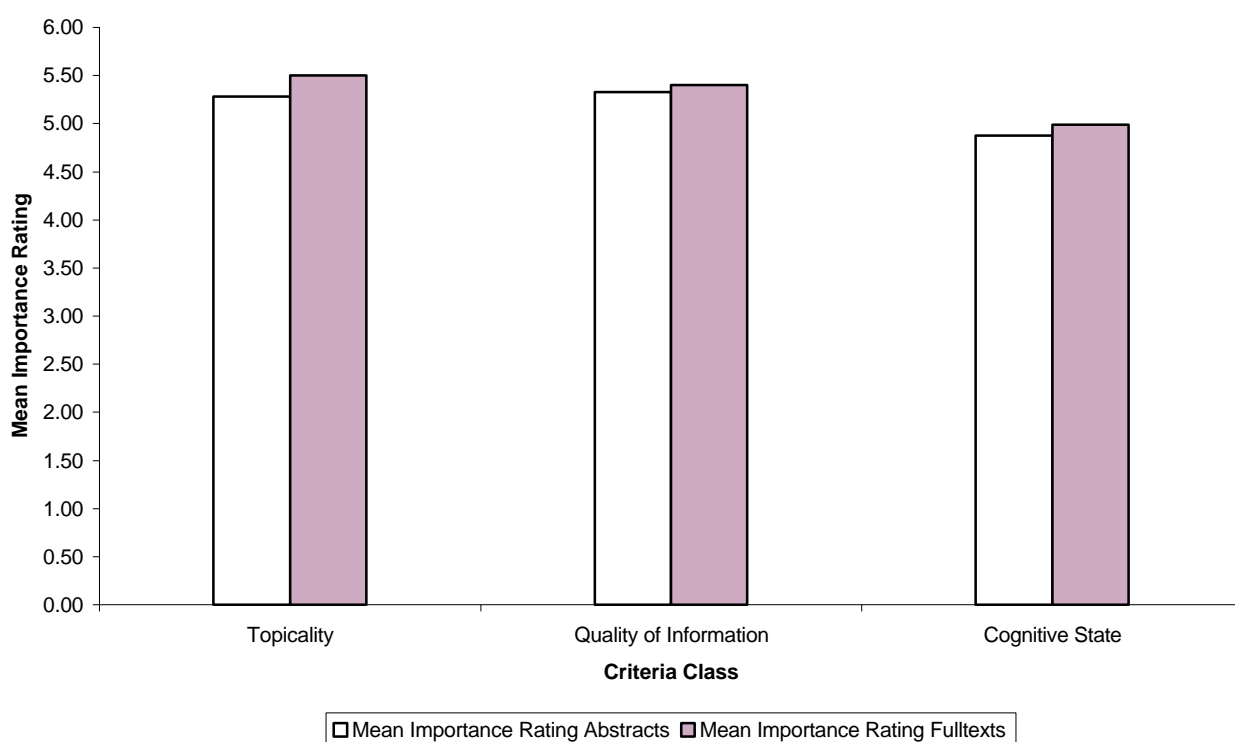


Figure 5.4. Mean Ratings of Criteria Classes Stage 1 and Stage 2

Quality of Information has the highest mean rating among the three at Stage 1, while *Topicality* becomes the most important category at Stage 2. The difference between the highest mean value and the lowest is 0.44 for Stage 1 and 0.51 for Stage 2, which suggests that the range of the differences is about a half point on a seven point scale. While *Cognitive State* remains the least important class of criteria for

both stages, it increased by about 0.11 at Stage 2. All three categories had increased ratings from Stage 1 to Stage 2. This finding may imply that the detailed information provided by full-texts allowed people to place higher demands on all three criteria classes.

The standard deviations for all three classes at both stages are very similar to one another. At Stage 1, *Topicality* (SD = 0.881) had the highest standard deviation. *Quality of Information* (SD = 0.686) had the lowest standard deviation. At Stage 2, *Cognitive State* (SD = 0.886) has the highest standard deviation. *Quality of Information* (SD = 0.710) continued to have the lowest standard deviation for Stage 2. Evidently, the participants tended to agree more with one another in their ratings on *Quality of Information* across the two stages. This may suggest that *Quality of Information* is somewhat different from dimensions of *Topicality* or *Cognitive State*. *Quality of Information* may be viewed a part of public knowledge that can be shared among participants collectively. *Cognitive State* and *Topicality*, as reflected in this study, can be viewed as a part of private knowledge, varying greatly by individuals.

Differences in the Ratings

The differences in the rating among categories between the two stages are illustrated in Figure 5.5. A positive value suggests that the average importance rating for a class of criteria increased from Stage 1 to Stage 2. Note that in this chart all three classes have positive values.

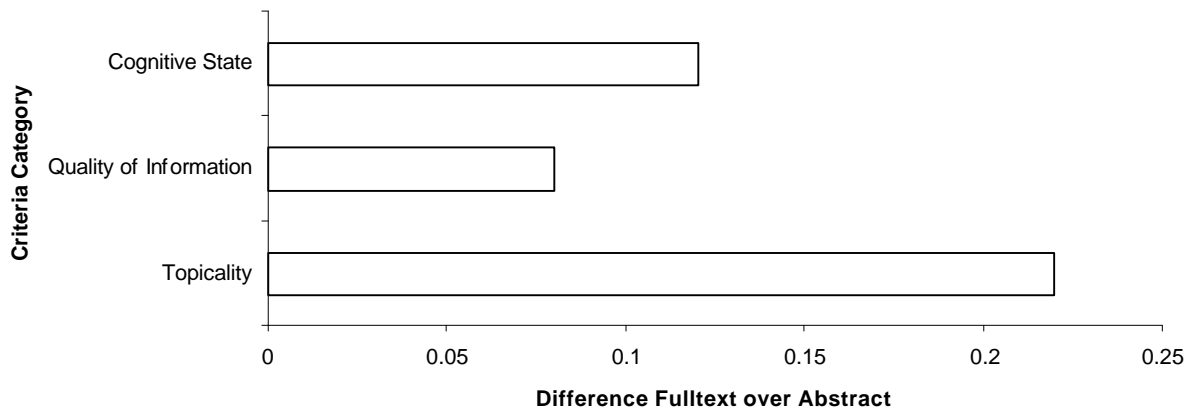


Figure 5.5. Difference in Mean Importance Ratings by Criteria Classes

The *Topicality* category has the greatest change in the importance ratings: it increased by 0.22, which is less than a quarter of one point on a seven-point scale. The average importance ratings for the category *Cognitive State* increased by 0.11, whereas *Quality of Information* had the least change, increasing at about 0.08 at Stage 2.

The stronger emphasis that participants put on *Topicality* at Stage 2 may have something to do with the nature of the composition task. At Stage 2, the participants were required to complete an outline. Their focus would be on how to collect facts and information from the full-texts concerning the topic of “The Year 2000 Problem and its Social Effect.” That is, the task was basically to collect facts, extract sentences and passages from the texts, and integrate important points of arguments reflected in a number of documents. Consequently, at Stage 2, whether or not the documents reviewed actually included information on the topic became crucial for the composition purpose. The value of *Topicality* was thus promoted to a greater extent than others at Stage 2.

Quality of Information as a group of criteria appears to be rather stable. While at Stage 1, it was rated as the most important class, at Stage 2 it dropped to the second. The relatively small change in the mean importance ratings of the two stages may suggest that quality is an issue that matters for undergraduate students across the stages, and consequently it did not change much.

Cognitive State was the class that had the lowest ratings for both stages, though still above 4.5 on a seven-point scale. Collectively speaking, the group increased by 0.11 at Stage 2. This suggests that, while paying attention to *Topicality* and *Quality of Information*, the participants' cognitive requirements for the documents also became stronger.

Reliability Tests for the a priori Model

Reliability tests were performed to examine the internal consistency of the elements in the researcher's classification scheme from a statistical standpoint. Reliability coefficient (Cronbach's alpha) were obtained based on correlations among criteria elements. Table 5.9 reports the reliability values of the three factors for Stage 1 and Stage 2. Most of the values are around 0.60, with *Quality of Information* holding the highest rates across the stages. The reliability value for *Cognitive State* at Stage 1 is the lowest among all, which is less than 0.50. Overall the test showed relatively low reliability for the researcher's a priori model.

Table 5.9
Reliability for Researcher's Model for Stage 1 and Stage 2

<i>Factors</i>	<i>Stage 1</i>	<i>Stage 2</i>
Factor 1 (Topicality)	.58	.62
Factor 2 (Quality of Information)	.67	.63
Factor 3 (Cognitive State)	.47	.62

Hotelling T² Test and Factor Analysis

Two statistical techniques were applied to examine the data pattern of criteria at the macro categorical level. First, the Hotelling T² technique was used to test the significance of a multivariate change between the stages. Following that, a Factor Analysis was performed to construct statistically valid clusters of criteria and to see if there was a correspondence between the suggested criteria classes and the factor classes. Both methods provide a statistical view of the patterns in the participants' ratings of criteria.

Hotelling T² Test

Based on the classification illustrated by Table 5.8, the Hotelling T² test resulted in a statistically significant ($T^2 = 11.012, p < 0.02$) difference between the two stages. This suggests that the global change of the multivariate structure is significant from Stage 1 to Stage 2. When examining the significance of individual sets of variables, it was found that among the three classes, *Topicality* is the only factor that has a statistically significant value. Neither *Quality of Information* nor

Cognitive State was significant when examined alone. Moreover, the single significant factor *Topicality* also became insignificant after a Bonferroni adjustment².

The test results demonstrate a statistically significant change in the overall ratings of the criteria sets. However, among the three individual groups, the change in *Topicality* contributes the most to the significance of the overall change. The participants' view of the importance of the criteria increased significantly on the *Topicality* dimension. This, as discussed earlier, may be related to the nature of the participants' writing task, which seems to promote strong demands for standards of *Topicality*.

Factor Analysis

Four-factor solutions. The purpose of Factor Analysis is to generate groupings according to the statistical distribution of the data. Based on the participants' importance ratings, a preliminary Factor Analysis produced two separate four-factor solutions, one for each stage. Table 5.10 presents the results of these four-factor solutions.

² Bonferroni adjustment serves as a conservative way for testing and guiding problems in multiple testing. The adjustment lowers the probability of making Type I Errors across the three comparisons.

Table 5.10
The Four Factor Solutions

<i>Factors</i>	<i>Stage 1</i>	<i>Stage 2</i>
Factor 1	Discuss Y2K and its social effect Information up to date Rich, well-rounded information Factual information and actual data Accuracy and trustworthiness Deepen my understanding	Provide definition of Y2K Information Up to date Rich, well-rounded information Factual information and actual data Accuracy and trustworthiness
Factor 2	Cover Y2K origin and causes Fresh and unique approach Consistent with previous knowledge Provide definition of Y2K	Cover Y2K origin and causes Fresh and unique approach Consistent with previous knowledge New information and new ideas Interesting and enjoyable
Factor 3	Understandable, not too technically complex Interesting and enjoyable Issues are real and important Clear and well-organized information	Deepen my understanding Discuss Y2K and its social effect Issues are real and important
Factor 4	New information and new ideas	Understandable, not too technically complex Clear and well-organized information

There are several common elements for the four factors at the two stages. For instance, Factor 1 of Stage 1 shares four common items with Factor 1 of Stage 2. Notice that for Stage 1, Factor 4 was formed by a single criterion “New information and new ideas;” whereas Factor 4 of Stage 2 is composed by “Understandable, not too technically complex” and “Clear and well-organized information.” The similarity of the two four-factor solutions (similarity = number of common elements/total number of criteria) is 53%.

Factor loadings for the 15 criteria and the percentages of variance explained by each factor provide a better idea of the structure of the four factors. The tables below display the factor loading values of each of the criteria under the four-factor solution structure for Stage 1 (Table 5.11) and Stage 2 (Table 5.12). The factor loadings were the results of Varimax method. Varimax is a loading criterion that rotates the axes of the coordinate correlation systems to produce factors solutions that are orthogonal.

Table 5.11
Factor Loadings for the Four-Factor Solution of Stage 1

<i>Factors</i>	<i>Stage 1</i>	<i>Factor Loading</i>	<i>Factor Variance Explained</i> (Total Variance Explained by the Solution: 61%)
Factor 1	Discuss Y2K and its social effect	.69	24%
	Information up to date	.76	
	Rich, well-rounded information	.79	
	Factual information and actual data	.71	
	Deepen my understanding	.64	
	Accuracy and trustworthiness	.76	
Factor 2	Cover Y2K origin and causes	.71	16%
	Fresh and unique approach	.65	
	Consistent with previous knowledge	.71	
	Provide definition of Y2K	.58	
Factor 3	Understandable, not too technically complex	.63	14%
	Interesting and enjoyable	.75	
	Issues are real and important	.71	
	Clear and well-organized information	.46	
Factor 4	New information and new ideas	.85	8%

At Stage 1 all of the elements in the four factors have relatively high loadings to their factors. Elements in Factor 1 all have strong loadings, among them, “Rich, well-rounded information” is the leading variable. For Factor 2, “Provide definition of Y2K” has the weakest loading while “Cover Y2K origin and causes” and “Consistent with previous knowledge” are the strongest. The leading variable for Factor 3 is “Interesting and enjoyable,” whereas “Clear and well-organized information” is the weakest. The single variable “New information and new ideas” has a very strong loading to establish an independent factor by itself. In terms of the overall model, the entire solution structure explains about 61% of the variance.

Specifically, Factor 1 explains about 24% of the variance, Factor 2 does 16% and Factor 3 does 14%. The final factor explains about 8% of the variance, and that indicates that Factor 4 is relatively weak.

Table 5.12 below includes the factor loading information for the four-factor solution at Stage 2 and the proportion of the variance explained by each factor.

Table 5.12
Factor Loadings for the Four-Factor Solution of Stage 2

<i>Factors</i>	<i>Stage 2</i>	<i>Factor Loading</i>	<i>Factor Variance Explained</i> (Total variance explained by the solution: 63%)
Factor 1	Information up to date	.63	19%
	Rich, well-rounded information	.65	
	Factual information and actual data	.72	
	Provide definition of Y2K	.53	
	Accuracy and trustworthiness	.80	
Factor 2	Interesting and enjoyable	.57	17%
	Fresh and unique approach	.72	
	New information and new ideas	.82	
	Consistent with previous knowledge	.73	
	Cover Y2K origin and causes	.48	
Factor 3	Deepen my understanding	.64	12%
	Discuss Y2K and its social effect	.68	
	Issues are real and important	.51	
Factor 4	Understandable, not too technically complex	.85	14%
	Clear and well-organized information	.89	

For Stage 2, the factor solution explains about 63% of the total variance. Specifically, Factor 1 explains 19% of the variance. Within this factor, “Accuracy and trustworthiness” has the strongest loading, whereas “Provide definition of Y2K” has the weakest. The criterion “New information and new ideas” loaded the strongest for Factor 2, while “Cover Y2K origin and causes” loaded the weakest. Factor 2 explains 26% of the variance. Variables in Factor 3 have relatively lower loadings, with “Issues are real and important” loaded at the lowest value. Factor 3 explains about 12% of the variance. The two criteria in Factor 4 have relatively high loadings, and the factor itself explains about 14% of the variance.

The four-factor solutions for both stages contain valuable information that is worth studying. First of all, it seems difficult to fit the four factors into the three dimensions described by researcher's a priori structure of classification. Consequently, the idea of making three-factor solutions emerged and the results of that will be discussed in the sections below. On the other hand, one way of characterizing these criteria factors may be to simply identify them as having either an objective orientation or subjective orientation. For instance, Factor 1 of both stages includes mostly objective criteria; whereas Factor 3 of Stage 1 solution and Factor 2 of Stage 2 solution consist mainly of subjective criteria. The issue of objective versus subjective criteria will be further explored in Chapter 6 and Chapter 7.

Three-factor solutions. Given the use of a three-category structure above (i.e., *Topicality, Quality of Information, Cognitive State*), another Factor Analysis was performed where the 15 criteria were forced into a three-factor structure. Table 5.13 displays the resulted three-factor solutions.

Table 5.13
The Three-Factor Solutions

<i>Factors</i>	Stage 1	Stage 2
Factor 1	Discuss Y2K and its social effect Information up to date Rich, well-rounded information Factual information and actual data Deepen my understanding Accuracy and trustworthiness	Cover Y2K origin and causes Information up to date Rich, well-rounded information Factual information and actual data Provide definition of Y2K Accuracy and trustworthiness
Factor 2	Cover Y2K origin and causes Fresh and unique approach New information and new ideas Consistent with previous knowledge Provide definition of Y2K	Interesting and enjoyable Fresh and unique approach New information and new ideas Consistent with previous knowledge
Factor 3	Understandable, not too technically complex Interesting and enjoyable Issues are real and important Clear and well-organized information	Deepen my understanding Understandable, not too technically complex Discuss Y2K and its social effect Issues are real and important Clear and well-organized information

The two three-factor solutions have some similarities with the previously produced four-factor structures. For Stage 1, the criterion “New information and new ideas,” previously the single element of Factor 4, now resides in Factor 2. For Stage 2, “Understandable, not too technically complex” and “Clear and well-organized information,” the previous constituents of Factor 4, now are grouped under Factor 3. In this three-factor structure, there are also similarities of the elements included for the same factors. Within Factor 1, Stage 1 and Stage 2 share four items, and within Factor 2, Stage 1 and Stage 2 share three items. The last factor shares three items between the two stages. In total, Stage 1 and Stage 2 share ten elements in common. The similarity between the groupings is about 67%.

Table 5.14 and Table 5.15 include the factor load values for the criteria and the percentage of variance explained by each of the three factors. Since the first factor remains the same for both three- or four- factor structures, they have the same value for the proportion of variance explained (35%). At Stage 1, the criterion “New information and new ideas” is added to Factor 2, hence the rate of that factor increases by about 1%. At Stage 2, two items are added to Factor 3, increasing the proportion of the variance explained by 5%.

Table 5.14
Factor Loadings for the Three-Factor Solution of Stage 1

<i>Factors</i>	Stage 1	Factor Loading	Factor Variance Explained (Total variance explained by the solution: 54%)
Factor 1	Discuss Y2K and its social effect Information up to date Rich, well-rounded information Factual information and actual data Deepen my understanding Accuracy and trustworthiness	.69 .76 .80 .71 .63 .76	23%
Factor 2	Cover Y2K origin and causes Fresh and unique approach New information and new ideas Consistent with previous knowledge Provide definition of Y2K	.73 .57 .52 .68 .69	16%
Factor 3	Understandable, not too technically complex Interesting and enjoyable Issues are real and important Clear and well-organized information	.69 .75 .68 .54	15%

Table 5.15
Factor Loadings for Three-Factor Solution of Stage 2

<i>Factors</i>	Stage 2	Factor Loading	Factor Variance Explained (Total variance explained by the solution: 54%)
Factor 1	Cover Y2K origin and causes	.53	19%
	Information up to date	.63	
	Rich, well-rounded information	.62	
	Factual information and actual data	.70	
	Provide definition of Y2K	.56	
	Accuracy and trustworthiness	.79	
Factor 2	Interesting and enjoyable	.56	18%
	Fresh and unique approach	.72	
	New information and new ideas	.81	
	Consistent with previous knowledge	.73	
Factor 3	Deepen my understanding	.53	17%
	Understandable, not too technically complex	.75	
	Discuss Y2K and its social effect	.54	
	Issues are real and important	.70	
	Clear and well-organized information	.68	

The results of the reliability test for the three-factor structures are shown in Table 5.16. Note that overall the reliability values are better than the ones in researcher's original model. Note also that among the three factors, Factor 1 holds the highest reliability for both Stage 1 and Stage 2.

Table 5.16
Reliability for Three-Factor Solutions at Stage 1 and Stage 2

<i>Factors</i>	<i>Stage 1</i>	<i>Stage 2</i>
Factor 1	.83	.75
Factor 2	.69	.71
Factor 3	.67	.74

It is interesting that at both stages “Information up to date,” “Rich, well-rounded information” and “Accuracy and trustworthiness” belong to the first cluster, with a strong loading (L = loading value) of 0.76, 0.71, 0.81 for Stage 1 and 0.63, 0.62, 0.79 for Stage 2. In other words, from a statistical point of view, these three particular criteria have the tendency of belonging to the category that I label *Topicality*. On the other hand, in her study, Bateman (1998) named the factor that contains “About my topic,” “Credible,” and “Accurate” as *Information Credibility*. Alternatively, *Information Credibility* may also serve as a label to characterize the first factor of the two stages in this study.

Bateman’s factor analysis also resulted loading the criterion “Current” to the factor she labeled *Information Quality*. However, the correlation was relatively weak, and Bateman (1998b) therefore suggests that this particular criterion should be categorized into a separate dimension *Information Currency*. In this study, “Information up to date” was grouped into *Quality of Information* in researcher’s a priori classification, but the factor analysis solution at both stages loaded the criterion on the factor *Topicality* according to my label, or *Information Credibility* according to Bateman’s model.

For Factor 2, the statistical solutions at both stages group “Fresh and unique approach” (stage1 $L = 0.57$, stage2 $L = 0.72$) into the cluster that I label *Cognitive State*. Lastly, “Understandable, not too technically complex” (stage1 $L = 0.69$; stage2 $L = 0.75$) is strongly patterned in the class *Quality of Information* for both stages. This is consistent with Bateman’s result: the criterion “Understandable” was loaded on the factor *Information Quality* in her study.

To a certain degree, the factor solution structures produced in this study appear to make sense in that “Information up to date,” “Rich, well-rounded information” and “Accuracy and trustworthiness” can be considered as describing various aspects of the topicality of a document whereas “Understandable, not too technically complex” is very much related to the readability of the document. It seems appropriate to consider “Understandability” as defining *Quality of Information*. In the same vein, the criterion “Fresh and unique approach” may be taken as having an impact on the participants’ *Cognitive State* because if an approach is viewed as fresh and unique, it would bring some elements that are new and intellectually interesting to the existing knowledge base.

Shifted criteria in the three-factor solutions. As indicated earlier, the factor structures for Stage 1 and Stage 2 are very similar, sharing 10 elements in common. Five criteria shift their factor memberships from Stage 1 to Stage 2. Both “Discuss Y2K and its social effect” and “Deepen my understanding” shift from Factor 1 at Stage 1 to Factor 3 at Stage 2. On the other hand, both “Cover Y2K origin and causes” and “Provide definition of Y2K” shift from Factor 2 at Stage 1 to Factor 1 at

Stage 2. “Interesting and enjoyable” moves from Factor 3 at Stage 1 to Factor 2 at Stage 1. The change in these five factors suggests the possibility that these criteria hold somewhat different meanings between the two stages, or that they are applied in different ways by the participants between the two stages.

In terms of the differences between the statistical factors and the a priori grouping structure that I applied, the Stage 2 three-factor solution holds a greater similarity to the researcher’s model than the Stage 1 solution does. The Stage 2 solution shares eight common elements in grouping with researcher’s structure, whereas Stage 1 shares five. The criteria that are in different groupings in Stage 2 solution and in researcher’s model include “Information up to date,” “Rich, well-rounded information,” “Accuracy and trustworthiness,” “Fresh and unique approach,” “Deepen my understanding,” “Understandable, not too technically complex,” and “Discuss Y2K and its social effect.” This could suggest that these criteria contain multiple connotations and may have been interpreted differently as the participants read them and rate them.

As pointed out in Chapter 4, it is important to keep in mind that a statistical solution does not necessarily carry theoretical significance. Rather, it is a reflection of the algebraic structure of the data distribution and is used here analytically to consider possible theoretical explanations for the empirical findings. In this light, some of the groupings seem to make sense conceptually, and others suggest possible modification or adjustment of the researcher’s original model. There is more about this in the discussion section below and in the final chapter of the dissertation.

Discussion

A number of issues emerged from the analysis of the experimental data. This discussion starts with the findings of the micro level analysis, followed by those of the macro level analysis. The contributions of the Hotelling T^2 test and Factor Analysis are also considered. Some possibilities are offered regarding ratings and change patterns of both the individual criteria and criteria categories.

At the micro, individual level, the criteria ratings at Stage 2 followed a very similar pattern to Stage 1. The four highest ranked criteria at Stage 1 are also the top four for Stage 2, differing only by their ranking status. Furthermore, the five lowest ranked criteria for Stage 1 are also the same for Stage 2, varying again only by the exact ranked positions. With this overall picture in mind, the following sections discuss a number of typical criteria.

The Highly Rated Criteria

The criterion “Understandable, not too technically complex” had the highest mean rating at Stage 1, and it ranked number 4 at Stage 2. The mean rating decreased by 0.13 at Stage 2. Given the nature of the topic, and given that over half of the participants were freshmen, it is quite reasonable that at the first stage, while they were reading abstracts, the participants’ initial concern was whether the items are written in a manner that is comprehensible to them: they wanted to exclude difficult documents that appear to be over-loaded with technical terms and programming jargon. At the second stage, once participants have read the actual document, understandability becomes less of a concern, although it remains as an important criterion.

One intriguing thing is that during the Factor Analysis, the solution structures for both stages grouped the criterion “Understandable, not too technically complex” into the category that centers on *Quality of Information*. Such a grouping may make sense. When people talk about an item as not being too scientifically complex, they may be referring to either the quality or characteristics of the information. Since the Year 2000 problem is an issue that is associated with computers and programming, “understandability” can be viewed as more of a matter of the information content and presentation quality, and less as a reflection of the reader’s knowledge state. On the other hand, the researcher’s original structuring was based on the view point that understandability has a stronger connection with an individual’s knowledge level and therefore belongs to the category “Cognitive State.” Both groupings can be seen as reasonable, depending on such matters as task and domain in addition to an individual state of knowledge. This analysis also shows that many criteria have rich connotations, and can be grouped into different categories depending on the interpretation of the people who use them. The more we understand about these connotations, the more likely that we can use such an understanding as inspiration for design. There are several other criteria that seem debatable regarding criteria group membership. These insights were used to create a revised classification of the 15 criteria, which will be presented at the end of this section.

At Stage 2, the highly ranked criteria include “Discuss Y2K and its social effect” and “Issues are real and important.” It appears that at this stage while the

participants were close to the point of extracting information from the articles for the presentation outline, what was viewed as important was whether the articles contained the exact information needed for participants' construction of the outline. The criterion "Issues are real and important" weighed more at Stage 2, although it was also ranked highly in Stage 1. The low standard deviation rates for this criterion at both stages indicate that the participants were consistent in rating "Issues are real and important." Evidently, at both stages all participants paid attention to the importance of the subject matter. Moreover, this finding coincides with the results reported earlier on the participants' comments regarding other criteria used. As the participants described these other criteria, several of them suggested looking for articles that cover the seriousness of the Year 2000 problem. The seriousness of the Y2K is somewhat related to the criterion "Issues are real and important." This evidence from another angle provides support for the high ratings of "Issues are real and important" at both stages.

The Lowest Rated Criteria

The lowest rated criterion for both Stage 1 and Stage 2 was "Consistent with my knowledge." The high standard deviation rates at both stages indicate that the participants varied in assessment of the importance of this criterion. Even with a slightly higher rating at Stage 2 ($d = 0.10$), it seems that the participants considered this particular criterion to be relatively unimportant. One possible explanation of such a consistently low rating is that the participants did not have confidence in

their own knowledge about the topic or that they were not certain about the meaning of the new information that they received while reading the documents.

Another criterion that was viewed as relatively unimportant was “Fresh and unique approach,” which was ranked as the second to the last for both stages.

Participants also did not think that “New information and new ideas” mattered too much to their selection decisions. Note that the standard deviation values for this criterion at both stages were high, suggesting high variation in the ratings.

Interestingly, the four-factor solution produced by the initial factor analysis put “New information and new ideas” as the single element for an independent factor (Factor 4) at Stage 1. This may indicate that at Stage 1 “New information and new ideas” carries a unique connotation that was strong enough to make the criterion independent from the rest of the three factors. On the other hand, the criterion “Fresh and unique approach” is clustered with the category *Cognitive State* for both stages. This may imply that the participants perceived such a criterion more from a sense of cognition than from a sense of information quality, or that there maybe something else that is being tapped by this grouping.

Given the nature of the topic, the participants’ education levels, and the fact that their final written task was a simple information gathering and extraction effort, it is understandable that participants were not very much interested in whether the article contains arguments that are consistent with their view points and whether the documents provide new information or have theoretical insights and are conceptually inspirational to them. Their main focus for Stage 1 was

whether, judging from an abstract, the full-text article would be comprehensible to them; and as they moved to Stage 2, they adjusted their attention toward whether the article contained the discussion of Y2K and its social effect and whether the article included real and important issues.

Stability of Criterion Items

As indicated earlier, the change in criteria ratings across the two stages is, in general, relatively small in scale. The most changed criteria were “Issues are real and important” ($d = 0.39$) in terms of mean ratings, and “Information up to date” ($d_r = -5$) in terms of rankings. Note that both criteria belong to *Quality of Information* in the researcher’s original model, whereas in the statistically generated solutions, those two all belong to *Topicality* for both stages.

The criterion “Deepen my understanding” had the second largest change both from the view of ratings ($d = 0.34$) and the view of rankings ($d_r = 4$). This particular criterion was ranked number 9 at Stage 1 and became number 5 at Stage 2. The participants apparently valued the criterion “The information contained in these documents deepen my understanding of the social effects of the Year 2000 problem” more after they read and selected full-text articles. According to the statistically generated three-factor solutions, this criterion was previously included in Factor 1 *Topicality* ($L = 0.63$) for Stage 1 and becomes an element of Factor 3 *Quality of Information* ($L = 0.53$) at Stage 2. Note that the loadings for both stages are relatively low.

Several criteria changed relatively little over the stages. In terms of the rating, the two most stable criteria include “Fresh and unique approach” ($d = 0.01$) and “Accuracy and trustworthiness” ($d = 0.02$). Notice that even though the two criteria are both stable from Stage 1 to Stage 2, “Accuracy and trustworthiness” was viewed as highly important--ranked number 2 at Stage 1 and number 3 at Stage 2, whereas “Fresh and unique approach” was viewed as relatively unimportant, being the second lowest criterion for both stages.

If measured by ranked positions, the most consistent criteria are “Provide definition of Y2K,” “New information and new ideas,” “Fresh and unique approach,” and “Consistent with previous knowledge.” All have the exact same ranking status from Stage 1 to Stage 2. Among these four, “Provide definition of Y2K” was the only criterion that was ranked relatively highly, as the sixth most important criterion. The remaining three criteria were ranked as the three lowest criteria across the two stages.

One thing that deserves some attention is that on the one hand, “Deepen my understanding” changed greatly as measured both by rating and ranking; on the other hand, another cognitively related criterion “Consistent with previous knowledge” was the most stable and most unimportant criterion in the rating and ranking. It appears that this group of participants placed high importance on comprehension and knowledge enrichment. The importance of “Deepen my understanding” was increased at Stage 2. On the other hand, due to the specific task they were performing and due to their education background, these

participants cared much less about whether the arguments presented in the documents were in agreement with their previous knowledge. Therefore, they continuously rated “Consistent with previous knowledge” as the criterion of lowest importance.

The Importance Rating of Criteria Categories

On the macro, categorical level, *Cognitive State* was rated as the least important class at both stages. The most important category for Stage 1 was *Quality of Information*, whereas for Stage 2 it was *Topicality*. In terms of changes, all three classes increased their strength of rating in moving from Stage 1 to Stage 2.

Among the three, *Topicality* increased the most — participants assigned more weight ($d = 0.22$) at Stage 2. *Cognitive State* increased by 0.11, while *Quality of Information* increased the least by less than 0.10.

Within the *Topicality* class, the most changed criteria include “Cover Y2K origin and causes” ($d = 0.33$), “Discuss Y2K and its social effect” ($d = 0.25$), and “Factual information and actual data” ($d = 0.24$). The most stable criterion is “Provide definition of Y2K.” Within the class *Quality of Information*, the most changed criteria are “Issues are real and important” ($d = 0.39$), “Rich and well-rounded information” ($d = 0.15$), and “Information up to date” ($d = -0.14$). The most stable criteria include “Fresh and unique approach” ($d = 0.01$) and “Accuracy and trustworthiness” ($d = 0.02$). *Quality of Information* contains both the most changed criterion and the most stable criterion among all 15 criteria. Criteria under *Cognitive State* all changed no less than 0.10. The most changed criterion is “Deepen my

understanding” ($d = 0.34$), and the most stable criterion is “Consistent with previous knowledge” ($d = 0.10$).

Change Patterns and the Process Model

The importance ratings of criteria categories for Stage 1 and Stage 2 and change patterns thus produced are interesting. The results do not agree entirely with what the Process Model suggests. As explicated in Chapter 3, the Process Model envisions that there are changes in the use of criteria as users move from Stage 1 to Stage 2. The Hotelling T^2 Test confirmed that the change for a global multivariate structure is significant. On the other hand, the process model also predicts that users’ emphasis on topicality would decrease at the second stage and participants’ attention would be more geared towards elements that relate to their personal knowledge structure or cognitive state. In this study, *Topicality* obtained a significantly higher rating at Stage 2 after participants had studied the full-text articles. *Cognitive State* also increased, but not as strongly as *Topicality*. What is more, according to the Hotelling T^2 result, the only statistically significantly changed set is *Topicality*. Thus, the findings here do not totally agree with predictions of the Process Model, and may reflect situational differences between the scholarly ideal of research over time and the real world of college students with a constrained time task.

Factor Solution

The results from factor analysis suggest that “Information up to date,” “Rich, well-rounded information,” and “Accuracy and trustworthiness” are properties of

the factor that I see as *Topicality*, whereas “Understandable, not too technically complex” contributes to *Quality of Information*. “Fresh and unique approach” is linked to *Cognitive State*. This grouping seems sensible and shows some similarity with the researcher’s original model. While I recognize that the statistically valid factor structure should not serve as a complete semantic substitution for the conceptual model generated based on literature and research, I think that the results of Factor Analysis provide several reasonable possibilities for groupings of the criteria. Table 5.17 presents the final version of classification for the 15 criteria, modified based on my analysis of the results of the Factor Analysis. Note that both “Information up to date” and “Accuracy and trustworthiness” were elements of *Quality of Information* in the researcher’s a priori classification, and now these categories fall under *Topicality*. “Understandable, not too technically complex” belongs to *Cognitive State* in the researcher’s original model and now falls under *Quality of Information*. “Fresh and unique approach” was a component criterion of *Quality of Information*, and now falls within *Cognitive State*.

Table 5.17
Final Classification Model of the 15 Criteria

<i>Categories</i>	<i>Group Number</i>	<i>Criterion Items</i>
Topicality	1	Discuss Y2K and its social effect
	1	Information up to date
	1	Rich, well-rounded information
	1	Provide definition of Y2K
	1	Factual information and actual data
	1	Cover Y2K origin and causes
	1	Accuracy and trustworthiness
Quality of Information	2	Understandable, not too technically complex
	2	Issues are real and important
	2	Clear and well-organized information
Cognitive State	3	Deepen my understanding
	3	Interesting and enjoyable
	3	New information and new ideas
	3	Consistent with previous knowledge
	3	Fresh and unique approach

Post hoc Analysis

A post hoc analysis was performed on the weights and the change in the criteria categories grouped according to the modified structure. Table 5.18 lists the mean importance rating for the three groups based on the new structure. The results of the original model are also included in the table for comparison.

Table 5.18
Mean Importance Rating by Criteria Classes (Post hoc)

Criteria Categories	Stage 1 (Post hoc)	Stage 2 (Post hoc)	Stage 1 (Original)	Stage 2 (Original)
Topicality	5.40	5.53	5.28	5.50
Quality of Information	5.62	5.73	5.32	5.40
Cognitive State	4.56	4.70	4.88	4.99

Table 5.18 shows that according to the new grouping, the highest rated category for both Stage 1 and Stage 2 is *Quality of Information*. *Cognitive State* is the lowest rated category among the three for both stages. The results from the original model show the discrepancy in the rating of *Topicality*, which was the second highest at Stage 1, but rose to be the highest rated category at Stage 2. In the post hoc model, all three classes also had positive changes in moving to Stage 2, and the class that has the largest increase from Stage 1 to Stage 2 is *Cognitive State* ($d = 0.14$). The analysis of the original model showed that *Topicality* is the most increased class. *Topicality* ($d = 0.13$) is the second in the post hoc data, and *Quality of Information* ($d = 0.11$) is the least changed category. The difference between the most and least changed categories is very small, only at 0.03. *Topicality* has almost the same change as *Cognitive State*. Figure 5.6 illustrates the change in criteria categories from Stage 1 to Stage 2.

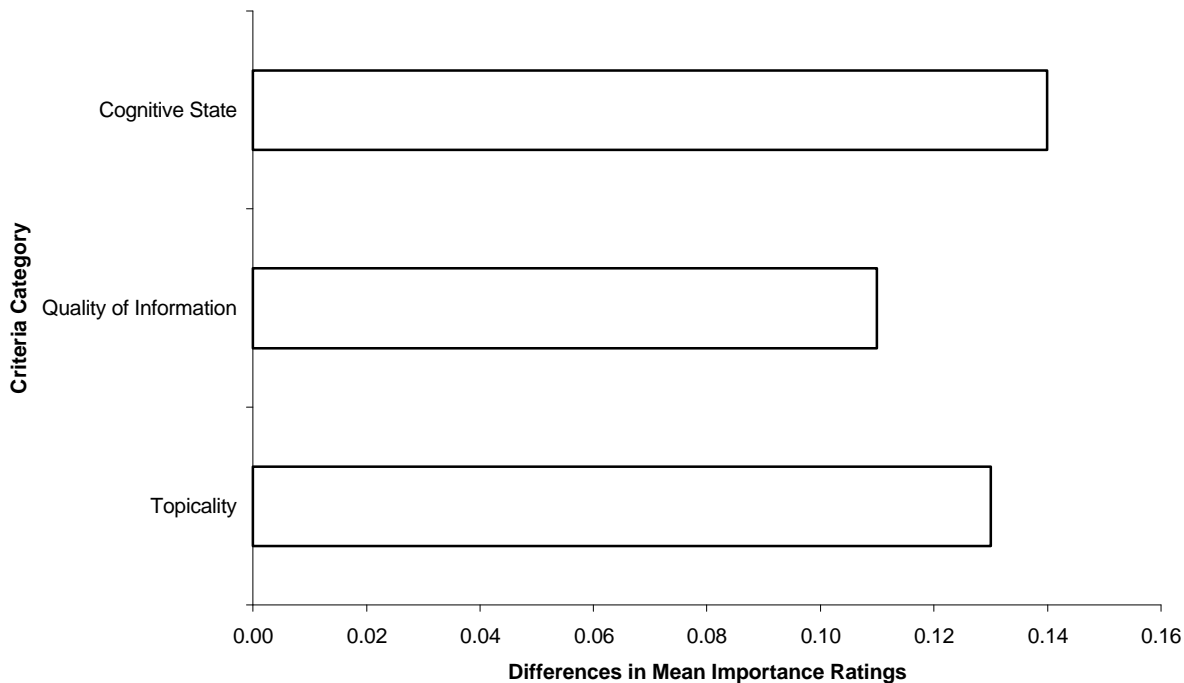


Figure 5.6. Differences in the Mean Importance Ratings (post hoc)

The post hoc analysis reveals that in terms of the importance rating, *Quality of Information* was rated the most important class for both stages, whereas *Cognitive State* was the least important category. However, in terms of the change in the ratings, *Cognitive State* increased the most, followed by *Topicality*, then by *Quality of Information*. The differences among the change values are too small to support further statements. Also note that the results of the post hoc only serve as an interesting way of data exploration. According to this exploration, *Cognitive State* increased the most at Stage 2, which provides some support for the Process Model that claims the dimension of cognitive state would be increasingly important during the full-text evaluation.

Summary

The experiment was designed in such a way that the participants had a simulated information need, and had to read abstracts and full-texts before they rated the importance of the 15 predefined criteria. Such a design challenges the assumption that the study of relevance criteria cannot be conducted in a controlled, laboratory setting. Both micro level and macro level analyses of the participants' importance ratings of criteria produced insightful results. The Hotelling T^2 test suggests that the change of the overall multivariate structure is significant, although the single significant factor is *Topicality*. Factor Analysis produced both four-factor solutions and three-factor solutions for Stage 1 and Stage 2. The three-factor solutions in combination with the factor loading values and proportions of variance explained provide a very good base for a modified factor structure, adjusted from the researcher's original classification model.

The results of Factor Analysis provide some indications with regard to whether the classification of criteria should be based purely on meanings of the criteria. It seems that relevance criteria cluster not necessarily along semantic attributes; they may convene according to their general nature. One observation was that some objective criteria tend to be in the same group of other objective criteria, the same can be said about subjective criteria. This leads to the idea of classifying criteria into a system of objective versus subjective. A detailed conceptual explication of a new taxonomy of criteria is presented in Chapter 6 and Chapter 7.

A number of issues concerning the nature of this project need to be brought to attention. Firstly, the data collected were participants' importance ratings of 15 criteria immediately following the reading of abstracts and then the reading of full-texts. The ratings reflect the participants' own perceptions of the importance of the criteria; they are NOT the participants' actual use of criteria. Secondly, the participants' task was a simple information extraction and integration type of composition, and the main purpose of assigning such a task to the participants was to stimulate a real information need for their document selection process. Therefore, by nature this kind of task should not be treated as a real life information seeking task that includes the writing of a research paper as an end product. The writing of a paper normally goes through an independent thought process, whereas for this project, the thinking that was involved was for the most part how to excerpt existing parts of the original documents and create an outline that combines the important points from several texts. Consequently, the criteria used in selecting a document for that purpose may be different from that used in a naturalistic research process. Thirdly, of the participants involved in this study, more than half were freshman students. This group may hold different perceptions of the importance of criteria from other people, for instance, people with advanced research experience.

In light of these three issues, and considering the different research mechanisms used for the two studies, it is necessary to note that the results of this laboratory experiment should not be directly compared to the results of the naturalistic study reported in next chapter. However, the two research efforts may

reveal overlaps and contrasts in phenomena that have conceptual implications for the study of users' criteria for relevance.

Chapter 6

RESULTS – THE NATURALISTIC STUDY

Introduction

The purpose of the naturalistic project is to study the use of relevance criteria in a situation that is as near as possible to a real information search process. Advanced Psychology graduate students were observed during their actual document selection processes. Their use of criteria was measured by their frequency of mentioning of criteria during the evaluation of bibliographic records and, then, during their evaluation of the full-text articles that were selected as relevant in the early stage. Data analysis was performed on two levels: a micro level analysis of frequency rates of individual criteria, and a macro level analysis of frequency rates of classes of criteria.

This chapter presents the results of the naturalistic project by first describing different types of the data collected. Next, a detailed report of the results is provided, starting with the micro level where the use of individual criteria was analyzed first on an individual participant basis, and then on a cross-participant basis. The macro level analysis begins by presenting the researcher's

reconceptualized working model with eight criteria classes for categorization. Under this working model, the focus of the analysis is not only on the use of the eight classes of criteria within each stage, but also on the evolving patterns of the criteria classes as participants move from Stage 1 to Stage 2 in a natural time-space. Next the participants' own perceptions of their use of criteria are summarized. Quotations from both the actual evaluation periods and the post evaluation interviews are provided to illustrate the participants' self-awareness of their use of criteria. At the end of the chapter, highlights of the naturalistic project and details of the researcher's interpretations of the findings of the naturalistic study are presented.

Characteristics of the Data

Participants

Ten PhD students from the Department of Psychology at the University of North Carolina at Chapel Hill (UNC-CH) agreed to participate in the study. One participant withdrew during the midst of the study, leaving a total number of nine participants. Six participants originally planned to conduct a meta-analysis for their topics, but only three indicated that they would carry out the original plan. The other three indicated that they had changed the nature of their papers. Instead of doing a meta-analysis, they completed a term paper, a literature review or a dissertation proposal.

A great majority of the participants were at an advanced stage of their PhD study when this research was conducted. Eight out of the nine people had either

completed their dissertation proposals or finished the preliminary writing of the proposal. One person completed her Master's thesis and was moving on to the PhD program. Of the nine participants, seven indicated that they had conducted experiments on their topics, or that experiments were in progress. Table 6.1 summarizes the academic status and research experience of the participants.

Table 6.1
Participants' Academic Status and Research Experience

Academic Standing	Number of Participants (<i>N</i> = 9)
Dissertation	1
Proposal in Progress	7
Passed Qualify Exam	1
Experiments on the topic completed or in progress	7

Types of Data

A variety of types of data were collected. The following description follows the order of the study process. The data for all nine cases were collected in a similar order:

- Pre-search Interview
- Record Evaluation
- Full-text Evaluation
- Post Document Evaluation Interview
- Participants' Scholarly Products

All the data summarized below were coded using the Nud*ist program (text analysis software).

Pre-search interview. The pre-search interviews were conducted in a semi-structured manner. Typically participants were asked to talk about the purpose of the search, the topic of interest, the types of articles they intended to look for, whether there were some authors known for the topic, and finally, what research or empirical experience they had on the topic.

All the interviews were transcribed, and then coded using Nud*ist.

Record evaluation. The evaluation of records normally started with the participants reading the title of a record, indicating a selection decision, and elaborating their reasons for that decision. This data serves as one of the two major sources of the data for the analysis reported in the later sections.

All of this verbal evaluation data was transcribed and the participants' comments for each record were coded by the criteria used and the selection decision made. For each record, each of the criteria mentioned was only coded once. A majority of the criteria included a positive and a negative value. For instance, if a participant stated that a record is not interesting, the record would be coded as "Interestingness" with the subcode "negative." Participants' general remarks about the use of some specific criteria were coded to a node called "general comments."

Full-text evaluation. There are two forms of evaluation data for the full-text articles. The first form includes the written comments that the participants wrote on document evaluation sheets after they read each article. These document evaluation sheets were provided for the participants to evaluate each of the articles they read. The data collected through document evaluation sheets included the participant's rating of the usefulness of the article being reviewed, and statements justify their

rating (Appendix H includes samples of the written evaluation). The second form of evaluation is the oral comments that participants made during the document evaluation interview. During that interview, all of the participants were instructed to go through each of the articles that they read. Participants first read the titles of the articles and, while referring to their written statements on the document evaluation sheets and looking at the actual articles, articulated their reasons for the ratings.

Both forms of evaluations were transcribed and data were coded under two separate nodes: “written” and “oral.” For two participants, there was a great deal of overlap between the written and oral comments; however, in the rest of the seven cases, oral comments were much more detailed and elaborated than the written ones. Each article reviewed was also coded by its usefulness rating. Many criteria contain both positive and negative values. Together with the record evaluation data, this data serves as the backbone for the data analysis reported in the later sections.

Post document evaluation interview. Nine separate semi-structured interviews were held after the participants had orally discussed the set of articles that they read. During the interviews, all of the participants were asked to define the concept “usefulness” and describe their interpretations of usefulness when they rated the articles. Next, participants were asked to characterize the collection of articles that they read by grouping the documents into categories. Following that, the researcher first explained the purpose of the study and then asked participants to reflect on their overall feelings of the use of criteria. The researchers asked the

participants whether in their opinions there were changes in the use of criteria, and whether new criteria emerged during the full-text evaluation. Participants were then asked to rank the criteria they used according to the importance of the criteria in the evaluation processes. At the end, the participants talked about the writing progress of the papers that they planned to compose for doing this project.

Participants' scholarly products. All of the nine participants initially intended to write a paper as a result of the literature search and reading articles. In the end, five participants indicated that they had completed a paper as scheduled. The other four said that there would be a delay in the completion of the paper. Currently the researcher has samples of the papers completed by three participants.

Data not Used in the Analysis

The dissertation study concentrates on the use of criteria, and therefore the analysis centers on the criteria mentioned by participants during Stage 1 and Stage 2. There are several sets of data that were processed but not analyzed. These include the data about participants' selection decisions when they reviewed bibliographic records and participants' usefulness ratings for each article they studied. Another set of data is the participants' research papers (other papers may be collected in the future). By looking at which articles the participants actually cited in their scholarly products and how the articles were cited, it is possible to map the entire document selection process. A comparison may be made between the bibliographic items selected at Stage 1 and the articles selected at Stage 2 and the articles cited in final papers at Stage 3. This would be an interesting complement to the present study, which may be pursued in the future.

Results

A basic measurement for the naturalistic project is the frequency of criteria used by participants, given the total number of documents reviewed. Specifically, the frequency count of a given criterion for a participant was obtained by counting the number of times that this criterion was mentioned by the participant. Each criterion was only counted once per document, even if the participant repeatedly referred to that criterion for one document. Positive and negative values of the criteria were also ignored for the time being. In general, the micro level analysis centered on the pattern of use for individual criteria, while the macro level analysis was intended to provide a broader, dimensional view of the use of the eight classes of criteria according to the revised model that is presented later in the chapter.

Micro Level Analysis

The micro level analysis looks at the use of individual criteria by participants on a case-by-case basis. The analysis then proceeds to consider the overall use of criteria across participants.

Use of Criteria by Participants

Search topics. All nine participants had relatively well-defined research topics before they came for the pre-search interview. Most of the participants knew some key authors in their areas of research, and all of them wanted to collect empirical articles or experimental reports. Some of them also wanted theoretical articles or review papers.

In terms of the search topics, six out of the nine participants had topics related to Social Psychology; the remaining three had topics that were related to Clinical Psychology.

Participant 1's topic was "self regulation failure and interpersonal relationship." He was interested in testing the applicability of a conceptual model called "Strength Model" in the domain of interpersonal relationships. He mentioned a prominent researcher (the originator of the model) when describing his topic. He then stated that he would be interested in getting experimental reports or theoretical papers.

Participant 2 was interested in the cognitive theory of pain. Specifically she focused on the effectiveness of the preemptive analgesia technique. She indicated that she became interested in the topic by reading an article that had an exemplary design with the comparison of effectiveness of pre- versus post- operative medications. She believed that experimental reports would be useful for her, and meanwhile, comparative studies of preemptive versus postoperative analgesia would be extremely helpful for her. She was also interested in review articles.

Participant 3 was looking for articles on the topic of "Stress and coping among young African American adolescents." She had already conducted empirical research (interviews and questionnaires) on the topic, and had completed the draft for her Master's thesis based on the research. Upon the suggestion of her committee, she decided to include two additional components in her thesis. One is the concept of reciprocal effect or reciprocal determinism; the other is the literature on the cultural environment of African American children. Consequently, the kinds

of articles that she was interested in were conceptual articles related to these matters.

Like Participant 3, Participant 4 had been working on an empirical project, which was his search topic. His research was part of a large scale, nationally funded project.

What we are interested in broadly is procedural justice, more specifically the interaction between lawyers, attorneys and their clients, and we are interested in knowing to what extent the trust that is developed between the client and his or her attorney affects that client's satisfaction with the outcome of that legal matter, above and beyond things like legal allocation, like how much money they can receive from the settlement.

Participant 4 had several specific authors in mind, and he indicated that he wanted to collect empirical studies on the topic of procedural justice in the legal domain. He also stated that both theoretical articles on procedural justice and review or descriptive articles on the interaction between attorneys and their clients would be useful for him.

Participant 5 had a unique task. He was asked to revise a submitted manuscript. The topic of the manuscript was "association between college students' first impressions of their instructors and students' class performance." He stated that, for the purpose of revising the manuscript, he wanted to search on four topics: a) first impressions of others' personalities (especially students' first impressions of instructors, and especially using the Big-Five approach), b) circumplex models (especially models of personality), c) Q-sort methods, and d) multidimensional scaling. As Participant 5 was relatively knowledgeable on the broad topic of

personality theories, he was familiar with the work of the major scholars in the field. He, thus, was particularly interested in several authors who were cited in the manuscript he was revising as well as an article title mentioned by one of his coauthors. Because he was interested in four topics, Participant 5 wanted a variety of types of articles: experimental reports, review papers, theoretical articles and method papers.

Participant 6 was interested in “individual group discontinuity.” He explains that this concept is about “the phenomenon or the tendency for individuals to be more cooperative than groups within the context of mixed motive situations.” He had worked with other people on two experimental projects on this topic. During the pre-search interview, he indicated that he intended to search for experimental studies by several authors that he had specified.

Participant 7 was interested in “cognitive therapy and Obsessive Compulsive Disorder (OCD).” She pointed out the difference between conventional cognitive behavioral therapy and cognitive therapy. She provided several researchers’ names, and stated that she was looking for experimental articles or review papers.

The topic that Participant 8 had was “intervention for women who had survived a cardiac event.” Specifically she was interested in psychosocial type of intervention in contrast with the conventional physical intervention such as exercise, diet, etc. She mentioned her project advisor as one of the scholars in this line of research. Participant 8 was looking for both experimental work and comparative studies of female versus male cardiac patients. She was also interested in review papers.

Participant 9 mentioned one representative author as he described his research topic. The topic was “self-esteem and people’s preferences of interaction partners.” He had published a study on the topic of self-esteem. He pointed out that there are two major theories of self-esteem in the current literature, and hence he wanted articles on these two theories as well as empirical studies.

Table 6.2 below provides an overview of the participants’ topics and the types of articles they were searching for.

Table 6.2
Participants' Search Topics

Participant Number	Topics	Know Key Authors	Type of Articles Preferred
Participant 1	Self regulation failure and interpersonal relationship	Yes	Experimental Reports; Theoretical Papers;
Participant 2	Preemptive analgesia and the cognitive theory of pain	No	Experimental Reports; Comparative Studies; Reviews
Participant 3	Stress and coping among young African American adolescents	Yes	Empirical Studies; Theoretical Articles
Participant 4	Procedural justice and the interaction between lawyers and their clients	Yes	Empirical Studies; Theoretical Articles
Participant 5	Association between college students' first impressions of their instructors and their class performance	Yes	Experimental Reports; Reviews; Theoretical Articles; Method Papers
Participant 6	Individual group discontinuity	Yes	Experimental Reports;
Participant 7	Cognitive therapy and Obsessive Compulsive Disorder (OCD)	Yes	Experimental Reports; Reviews
Participant 8	Intervention for women who had coronary heart disease	Yes	Experimental Reports; Comparative Studies; Reviews
Participant 9	Self-esteem and people's preferences of interaction partners	Yes	Experimental Reports; Theoretical Papers

Number of items reviewed. The number of items reviewed differs by the participants and by the stages. At Stage 1, the number of records reviewed ranged from a low of 46 to a high of 202; at Stage 2, the number of full-texts reviewed ranged from a low of 7 to a high of 39. Table 6.3 below lists the number of documents reviewed by the nine participants at the two stages. The average number of documents reviewed for Stage 1 is 84, and for Stage 2 is 27.

Table 6.3
Number of Documents Evaluated by Participants at Stage 1 and Stage 2

Participant Number	Number of Documents Reviewed Stage 1	Number of Documents Reviewed Stage 2
Participant 1	71	15
Participant 2	60	25
Participant 3	46	30
Participant 4	97	34
Participant 5	202	37
Participant 6	46	7
Participant 7	78	36
Participant 8	51	23
Participant 9	102	39

Note that Participant 6 only reviewed seven full-text articles, and according to him, that is partially because there was limited research on the topic and he already had some of the relevant articles.

Number of criteria employed by participants. The total number of criteria used varied both by the participants and by the stages. For definitions and examples of each of the criteria, refer to Appendix I. Table 6.4 lists the number of criteria each participant employed at different stages of document evaluation. The number of criteria used ranges from a low of 10 to a high of 37. The average

number of criteria mentioned is 23 for Stage 1, 20 for Stage 2 written, and 25 for Stage 2 oral. Figure 6.1 reports the number of criteria used by participants in a graphic display, with Stage 2 averaging the number of Stage 2 written and oral.

Table 6.4

Total Number of Criteria Used by Participants at Stage 1 and Stage 2

Participant Number	Stage 1	Stage 2 written	Stage 2 oral
Participant 1	28	17	29
Participant 2	32	24	37
Participant 3	14	24	27
Participant 4	14	10	12
Participant 5	37	27	35
Participant 6	16	12	13
Participant 7	13	28	29
Participant 8	23	17	22
Participant 9	27	17	22

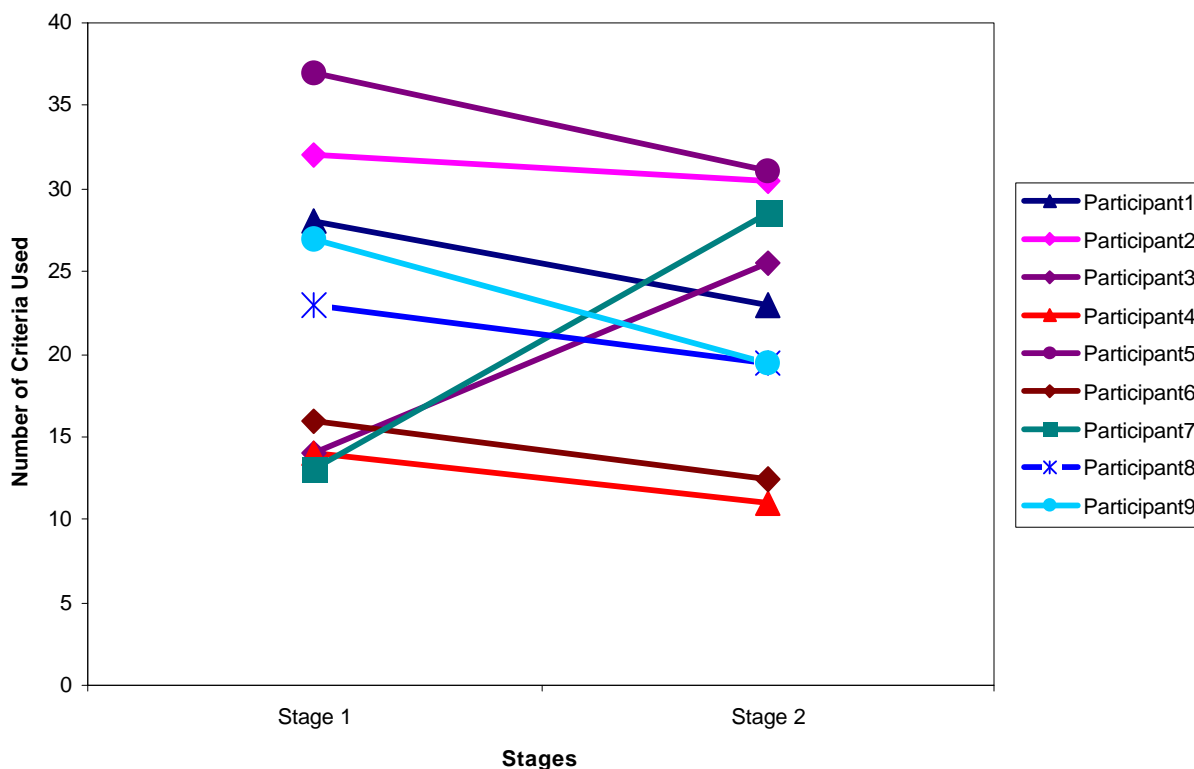


Figure 6.1. Number of Criteria Used by the Participants over the Stages

The number of criteria used by seven of the participants decreased as the participants moved from Stage 1 to Stage 2 written, but increased when they were prompted during Stage 2 oral. Only two participants, 3 and 7, showed a deviation from this pattern: the number of the criteria used increased from Stage 1 to Stage 2 written, it decreased or increased a little from Stage 2 written to Stage 2 oral.

It is possible that the decrease in the number of criteria employed from Stage 1 to Stage 2 written indicates a more focused use of criteria, while the increase suggests an expansion. A general observation of why many criteria were not used at Stage 2 is that they were related to the physical characteristics of the document such as “Author,” “Journal,” “Recency,” or “Language.” At Stage 2, when the participants examined the documents that had been selected using such criteria as

“Journal” or “Recency,” it is possible that the participants no longer needed to apply the criteria again. That might explain why the number of criteria was reduced for most of the participants. On the other hand, contrary to most of the participants, Participants 3 and 7 demonstrated an elaborated use of criteria for the full-text evaluation. Most of the criteria that they used that were unique to Stage 2 were related either to their cognitive needs, or to the usefulness of the documents to the experimental projects that the participants were conducting, or to the quality of the documents.

Actual use of criteria by the participants. For all nine participants, the specific use of criteria, as measured by relative frequency rates, varied from case to case. For each individual, the relative frequencies of the criteria employed at Stage 1 were related to those of Stage 2. Since Stage 2 contains both written and oral comments, the average of the relative frequency of Stage 2 written and Stage 2 oral was taken as the representative measurement for that stage. After excluding the unique criteria of a single stage, a contrast was made by listing the relative frequency of Stage 1 in comparison with that of Stage 2.

To reduce the size of this report, the following section displays the frequency table and chart of one participant. Among the nine participants, Participant 5 provided relatively elaborated use of criteria at both stages, and therefore his case was selected to illustrate the use of individual criteria at Stage 1 and 2 (see Table 6.5).

Table 6.5
Use of Individual Criteria by Participant 5

Criterion Item	Relative Frequency Stage 1 (N=202)	Average Relative Frequency Stage 2 (N=37)
Topical Focus	16%	31%
Topical Relatedness	46%	21%
Design	0%	34%
Method	6%	8%
Population	8%	10%
Nature of Study	7%	12%
Variables and Constructs	17%	7%
Clarity and Well-Written	1%	7%
Importance	2%	5%
Quality and Value	1%	10%
Scope	15%	6%
Type of Article	2%	8%
Author	17%	4%
Classic Study	2%	3%
Length of Article	1%	3%
Reference	1%	3%
Interestingness	13%	18%
Usefulness	2%	10%
Similar to What I do	4%	35%

Participant 5 reviewed a total of 202 records, and read a total of 37 full-text articles. At Stage 1, “Topical Relatedness” was mentioned most frequently (46%). Other frequently mentioned criteria include: “Author,” “Variables and Constructs,” “Topical Focus,” “Scope,” and “Interestingness.” At Stage 2, the frequently used criteria were “Similar to What I do,” “Design,” “Topical Focus,” and “Topical Relatedness.” Notice that “Design” was also used once at Stage 1, but in comparison to a total of 202 records, its relative frequency is 0% for that stage.

Figure 6.2 illustrates the content of the Table 6.5 in a graphic format. Criteria were sorted in a descending order by the average relative frequencies of Stage 2 for analysis purposes.

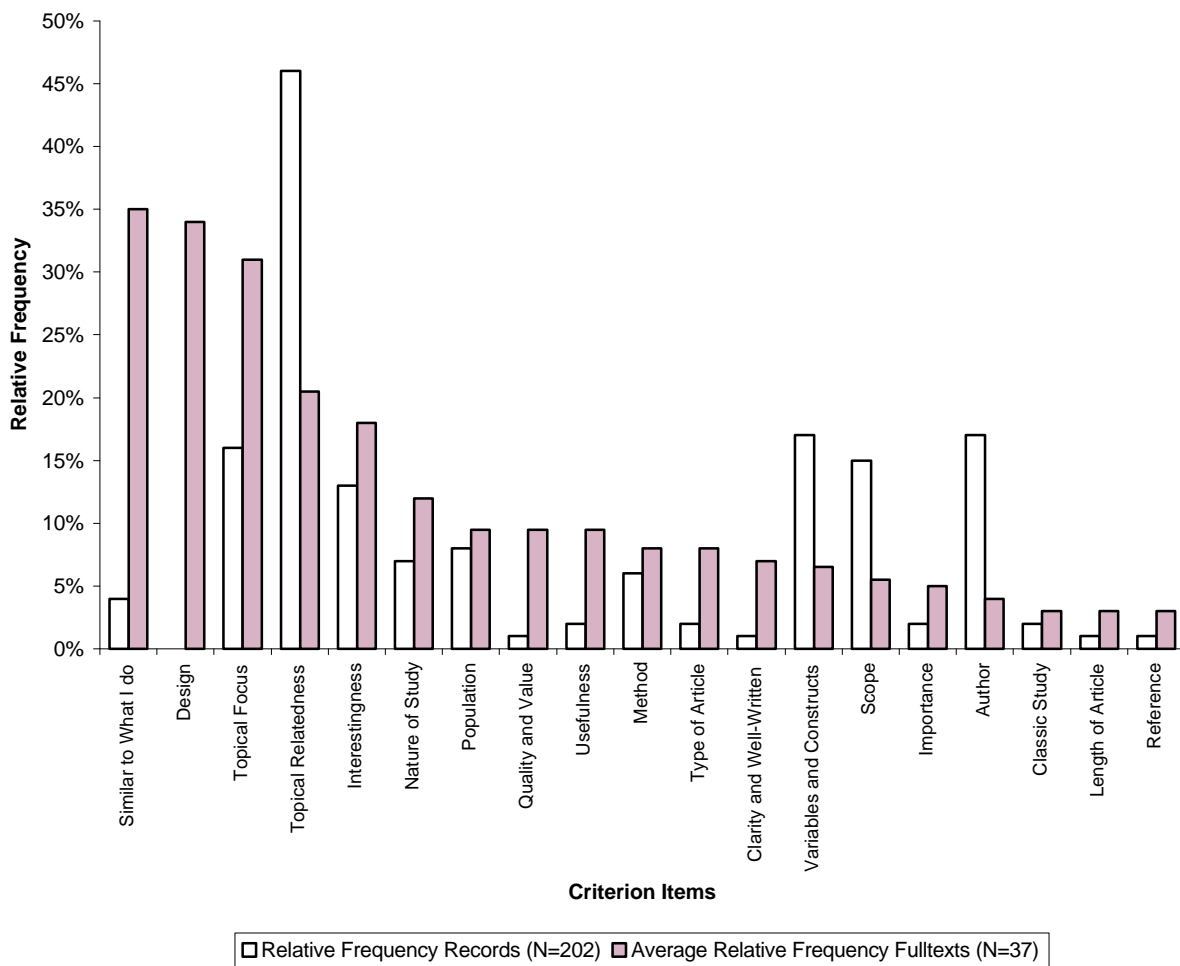


Figure 6.2. Use of Criteria by Participant 5 Stage 1 and Stage 2

Figure 6.2 shows that for Participant 5, the criteria that had changed a great deal from Stage 1 to Stage 2 include “Design” (increased Stage 2), “Similar to What I Do” (increased Stage 2), “Topical Focus” (increased Stage 2), “Topical Relatedness” (decreased Stage 2), and “Author” (decreased Stage 2). This suggests that at Stage 2, Participant 5 was much more focused on the details of research mechanisms, and the connection of the research reported in a full-text article to his own research

project. On the other hand, while the decrease in “Topical Relatedness” suggests that a general level of examination on topicality became much less significant at Stage 2; the increase of “Topical Focus” signifies a stronger emphasis on the much more specific evaluation of topicality at Stage 2. The criterion “Author” was used much less often at Stage 2, possibly because that it was applied frequently at Stage 1 as a criterion to include/exclude items and hence at Stage 2 it was no longer an essential criterion for accepting/rejecting a full-text article. Thus, while “Author” was mentioned at Stage 2, it was in the context of describing the author’s contributions to the participant’s area of study.

As evidence of how the use of criteria varied greatly from case to case, Participant 1 used the criteria “Topical Relatedness” and “Author” more frequently at Stage 2. For Participant 1, most of the criteria he mentioned increased from Stage 1 to 2. These criteria included “Usefulness,” “Topical Relatedness,” “Interestingness,” “Author,” “Topical Focus,” “Newness,” “Inspirational,” and “Importance,” among others. Only four criteria had decreased frequencies at Stage 2, and they were “Journal,” “Read Before,” “Domain,” and “Article Type.”

During the pre-search interview, Participant 1 indicated that his main purpose was to become more knowledgeable on the theory of “self regulation.” In addition, he was searching for some new design ideas. Since Participant 1 considered himself at the stage of expanding his knowledge base and collecting new ideas, he appeared to be increasingly interested in every aspect of a given study. This may explain why not only the uses of criteria such as “Newness,” “Interestingness,” and “Inspirational” increased, but also “Topical Relatedness” and

“Author” were mentioned more frequently at Stage 2. It is worth noting that during the post evaluation interview, Participant 1 claimed that he felt that at Stage 2 there was a reduction in the importance of criteria such as “Author” and “Journal.” However, in his actual use, “Author” increased at Stage 2 contrary to his reflection while “Journal” was decreased at Stage 2 as he had thought that it would.

In Participant 2’s case, the frequencies of most of the common criteria decreased at Stage 2. These decreased criteria include “Interestingness,” “Importance,” “Article Type,” “Results,” “Topical Focus,” “Topical Relatedness,” and “Design,” among others. Participant 2 had a well-defined idea of what she was going to do, and at Stage 1 she examined very closely the various elements of topicality, research structure (design, population, results, etc.), interestingness, etc. At Stage 2, these elements still had high frequencies, but compared to Stage 1, they were all reduced. The few increasingly used criteria include “Techniques,” “Quality and Value,” and “Well-written.” Participant 2 seemed to pay more attention to the techniques applied for pre- versus post-operative analgesia. Meanwhile, she considered very much about the writing style of some of the animal studies she read, and consequently, she used the criteria “Quality and Value” and “Well-Written” more often at Stage 2.

Participant 3 wanted to gather articles in order to add two sections to her thesis. She applied a larger number of criteria at Stage 2 than at Stage 1. That is, during the full-text evaluation, she incorporated more considerations into her reasoning. In terms of frequency for criteria commonly used for the two stages, those criteria that were used less frequently during Stage 2 include “Population”

and “Evidence of Effect.” Criteria that increased in use to some extent at Stage 2 were “Data,” “Quality and Value,” and “Nature of the Study.” Several criteria had a minor increase in frequency, and others did not change from Stage 1 to Stage 2.

Participant 3 paid much attention to “Population” at Stage 1. At Stage 1, she looked mainly at the “age range” and “ethnicity” of the subjects of a given study, and selected the items that included a population appropriate to her interests. At Stage 2, her attention was more on whether an article contained data, the quality of the paper, and the nature of the research reported.

Most of the criteria used by Participant 4 had an increased frequency rate at Stage 2. The greatly increased criteria include “Topical Focus,” “Topical Relatedness,” “Domain,” “Link to My Study,” and “Concepts.” Only one criterion was used less often at Stage 2, and that criterion was “Is About.” At Stage 1, since about one third of the records had no abstracts, Participant 4 went to some effort to figure out what a given document was about. At Stage 2 with the full-text articles, the participant was able to make quicker judgments on topicality as well as the connection between the study reported and his own project by utilizing the typical structure of psychological research reports to locate information needed for assessment purposes.

The criteria that Participant 6 used at Stage 1 were very different from what he used for Stage 2 evaluation. At Stage 1, the frequently used criteria included “Topical Relatedness,” “Nature of the Study,” and “Author.” At Stage 2 using seven articles, he consistently applied criteria such as “Topical Focus,” “Variables and Constructs,” “Statistics Analysis.” All four common criteria had increased

frequencies at Stage 2. Among them, “Suitable for Meta-analysis” increased the most, while “Interestingness” increased the second. “Data” was used more often at Stage 2, and “Quality and Value” also increased to some extent, but its increase was the least.

Among all nine participants, Participant 6 had the fewest number of full-texts to read at Stage 2. Yet compared to Stage 1, he provided a relatively detailed rationale for evaluation. During the post evaluation interview, he stated that he employed three criteria for his full-text evaluation: the article should a) focus on the topic of his interest; b) include the appropriate construct and have the right dependent variables; and c) contain detailed statistical information, allowing the extraction of effect size for meta-analysis purposes. These three criteria were not mentioned at Stage 1, and the participant’s evaluations at Stage 1 were relatively brief and were centered on “Topical Relatedness” and “Nature of the Study.” Consequently, most of the criteria that he mentioned are not comparable across the two stages.

Participant 7 pointed out in the pre-search interview that the purpose of her search was to collect articles in preparing for her dissertation proposal. She had a basic idea of what she was going to do for her dissertation research, and she had conducted a project on a similar subject. Recall that Participant 7 and Participant 3 were the only two who applied a greater number of criteria at Stage 2 than they did at Stage 1. Participant 7 was very focused during the record evaluation, using a constant set of criteria. At Stage 2, while emphasizing the connection between the

work reported and her own research, she employed a finer reasoning scheme to differentiate the documents read.

The actual use of criteria was measured through the frequency of the criteria commonly used for the two stages. In that regard, the frequently used criteria by Participant 7 at Stage 1 included “Topical Focus,” “Author,” and “Level.” At Stage 2, the frequently used criteria became “Link to My Study,” “Treatment,” “Usefulness,” “Nature of the Study,” “Newness,” “Quality and Value,” and “Techniques.” Among the seven commonly employed criteria, “Level” was the only one that was used less often at Stage 2. Other criteria such as “Nature of the Study,” “Interestingness,” “Population,” “Design,” and “Domain” were applied more often during Stage 2. “Similar to What I Do” had a minor increase at Stage 2.

It appears that Participant 7 used “Level” more often at Stage 1 to eliminate items that seemed too elementary or too introductory to her. However, at Stage 2, not only did she form a finer evaluation scheme, but she also demonstrated an increased interest in linking the study reported to her own research. She also focused on the mechanism and quality of the study, as well as the conceptual and methodological newness of the study. All of these were related to her need of selecting papers that would help constructing her dissertation proposal. It appears that the availability of the full-text allowed her to use the criteria that considered information that was not usually part of a bibliographic representation.

For the full-text evaluation, Participant 8 employed “Design” much more frequently than for the record evaluation. Other increasingly used criteria include “Usefulness” and “Justification of My Study.” A largely decreased criterion was

“Interestingness,” dropping about 35% from Stage 1. Apparently at Stage 1, Participant 8 applied “Interestingness” as a criterion to exclude uninteresting items. At Stage 2, since all the documents selected fulfill the “Interestingness” requirement, she concentrated more on the design of a study. At Stage 1, the frequently used criteria were “Topical Focus,” “Interestingness,” “Article Type,” and “Variables and Constructs.” At Stage 2, the focal criteria included “Design,” “Topical Focus,” “Usefulness,” “Nature of the Study,” and “Justification of My Study.”

It is interesting that Participant 8 paid special attention to the criterion “Justification of My Study” for both stages and even more so for Stage 2 than 1. According to Participant 8, there is much research effort on men with coronary heart disease, and the intervention literature is primarily oriented to men. Consequently, as she searched through the literature, she looked specifically for articles that compared men and women or contained certain statements about the need for more research on female cardiac patients. These documents were useful to her in that they provide a good justification for her research, which was centering on the psychosocial interventions for women who have experienced a cardiac event.

Participant 9 had a rather focused idea for his topic (self-esteem and people’s preferences of interaction partners) before he started the search. In his case, the frequently used criteria at Stage 1 were “Topical Focus,” “Topical Relatedness,” “Similar to What I Do,” “Interestingness,” and “Affective.” At Stage 2, “Topical Focus” remained to be the most frequently used criteria; other frequently mentioned criteria included “Usefulness,” “Nature of the Study,” and “Author.” Criteria that had increased frequencies at Stage 2 were “Topical Focus,” “Usefulness,” “Theory,”

and “Author.” Criteria that were used less often at Stage 2 included “Topical Relatedness,” “Interestingness,” “Affective,” and “Journal.” According to Participant 9, there were three major criteria that he used for Stage 2 evaluation. The articles must be: a) topically focused on people choosing an interaction partner and the two theories of self-esteem; b) involving certain variables; c) items written by a key author. While the first criterion explains the increasing frequency for both “Topical Focus” and “Theory,” the third criterion justifies why “Author” was also used more often at Stage 2.

In summary, the use of criteria and the change in the use of criteria across the two stages varied greatly by participants. However, as explained in previous paragraphs, for most of the cases the use patterns are justifiable in consideration of the participants’ specific research interests and needs for information. The changes from Stage 1 to 2 also seem to reflect differences in the nature of bibliographic records versus full-text documents. There was less of a learning effect in moving from Stage 1 to 2 than in the pilot study for this dissertation (Tang & Solomon, 1998) possibly because of the advanced level of the participants here.

Use of Criteria across Participants

A general sense of the use of criteria across the participants was acquired by adding the raw frequency counts of each of the criteria and normalizing them by the total number of items reviewed by the nine participants. The total relative frequency rates of a given criteria for the two stages was thus established, with Stage 2 averaging the total relative frequencies of written and oral. The total relative frequencies of both stages were compared while excluding the criteria that

were unique to a single stage. The criteria that were unique to Stage 1 include “Title Indicativeness,” “Only Title Available,” “Language,” “Geographic Location,” and “Familiarity,” among others. The criteria that were unique to Stage 2 include “Treatment,” “Support My View,” “Statistical Analysis,” “Sophistication,” “Interpretation,” and “Author Bias,” among others. Table 6.6 lists the total relative frequencies of both stages.

Table 6.6
Total Relative Frequency of Criteria Use Stage 1 and Stage 2

Criteria	Total Relative Frequency Stage 1 (N=753)	Average Total Relative Frequency Stage 2 (N=246)
Topical Relatedness	38%	27%
Topical Focus	26%	32%
Interestingness	16%	14%
Author	11%	6%
Population	8%	8%
Variables	8%	9%
Journal	7%	1%
Is About	7%	2%
Technique	6%	8%
Nature of the Study	6%	11%
Scope	6%	4%
Type of Article	6%	8%
Design	6%	17%
Publication Date	5%	2%
Domain	4%	6%
Importance	4%	6%
Quality and Value	3%	11%
Similar to What I Do	3%	8%
Read Before	3%	1%
Link to My Study	3%	11%
Results	3%	12%
Method	2%	4%
Helpfulness	2%	2%
Affective	2%	3%
Procedure	2%	9%
Newness	1%	6%
Reference	1%	4%
Usefulness	1%	23%
Classic Study	1%	3%
Theory	1%	6%
Understandability	1%	1%
Level	1%	1%
Data	1%	6%
Background Information	1%	1%
Article Length	1%	1%
Concepts	1%	3%
Justification of My Study	1%	3%
Readability	1%	1%
Evidence of Effect	1%	2%

The most frequently used criteria (freq. = relative frequency of a given criterion) at Stage 1 across the participants were “Topical Relatedness” (freq. = 38%), “Topical Focus” (freq. = 26%), “Interestingness” (freq. = 16%), and “Author” (freq. = 11%). At Stage 2, the most frequently used criteria included “Topical Focus” (freq. = 32%), “Topical Relatedness” (freq. = 27%), “Usefulness” (freq. = 23%), “Design” (freq. = 17%), “Interestingness” (freq. = 14%), “Results” (freq. = 12%), “Nature of the Study” (freq. = 11%), “Link to My Study” (freq. = 11%), and “Quality and Value” (freq. = 11%).

Figure 6.3 below provides a visualization of the data patterns as demonstrated in Table 6.6, with the distribution sorted by the average total relative frequencies in Stage 2.

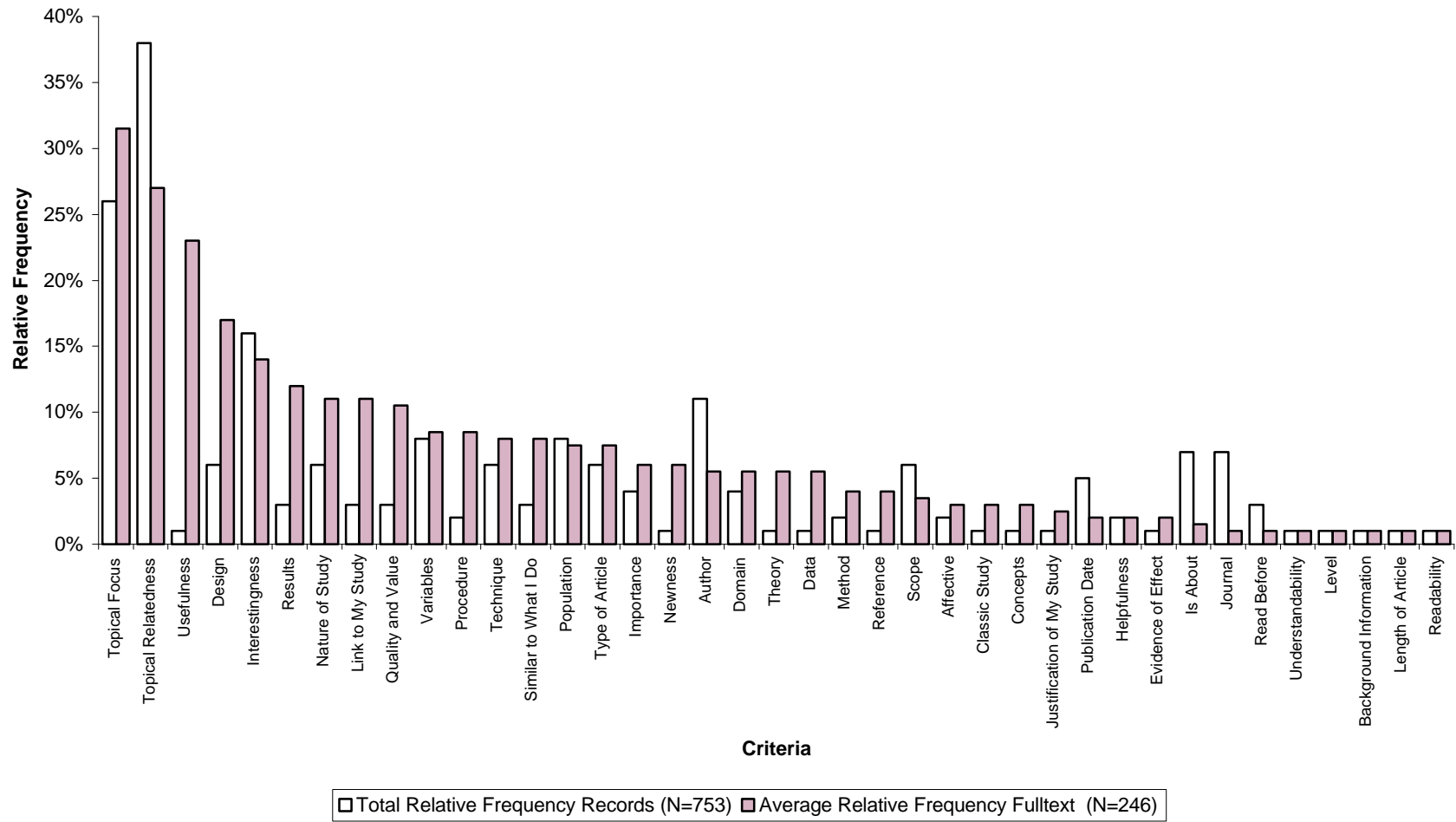


Figure 6.3. Use of Criteria across Participants Stage 1 and Stage 2

The criteria that had reduced frequency rates (d = average total relative frequency Stage 2 – total relative frequency Stage 1) from Stage 1 to Stage 2 were “Topical Relatedness” (d = -11%), “Journal” (d = -6%), “Is About” (d = -5%) and “Author” (d = -5%), among others. Many criteria had increased frequency rates at Stage 2. Among these were “Usefulness” (d = 22%), “Design” (d = 11%), “Results” (d = 9%), “Quality and Value” (d = 8%), and “Link to My Study” (d = 8%). Figure 6.4 provides another view of the use of criteria across participants, emphasizing change between the two stages.

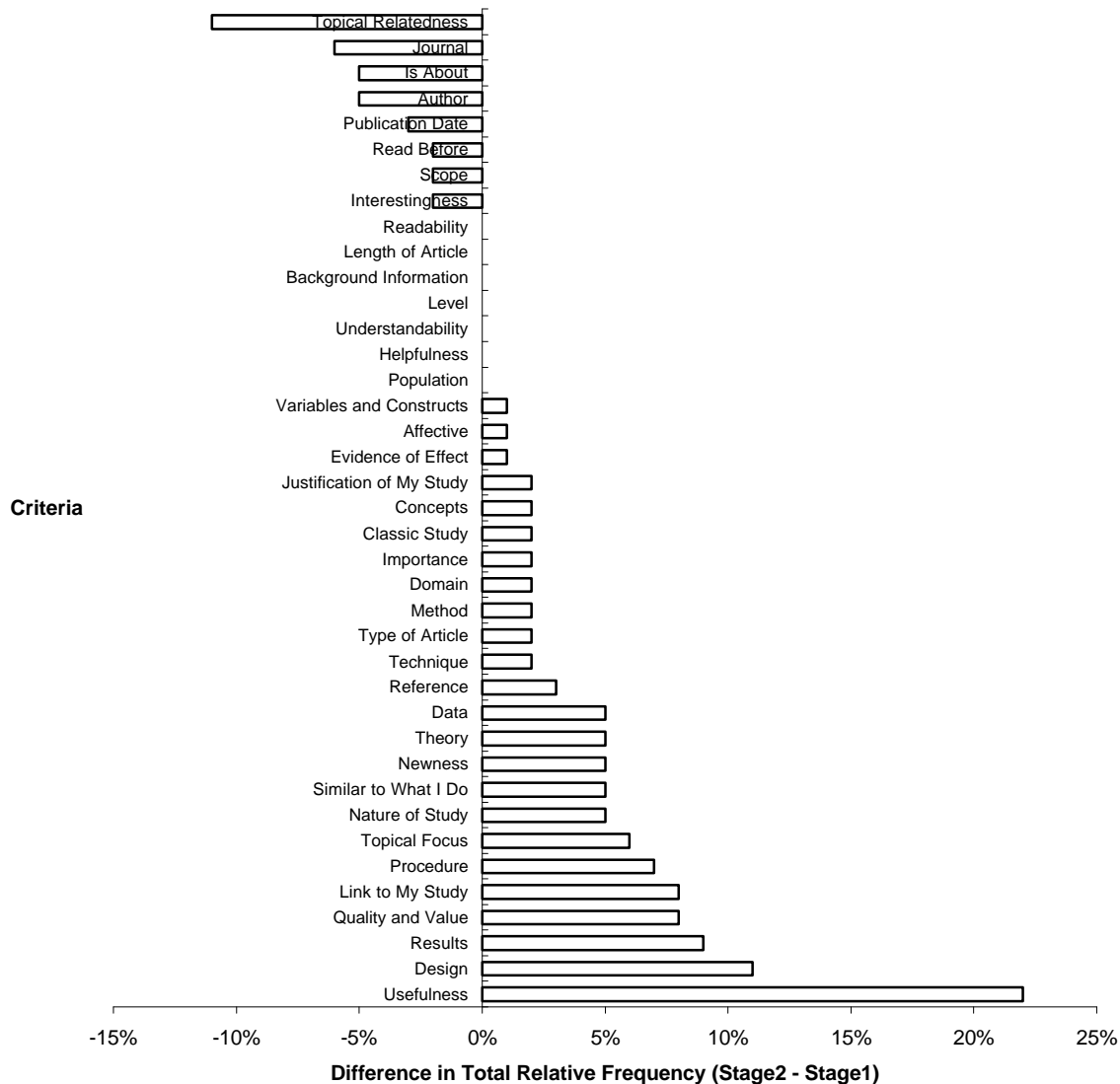


Figure 6.4. Differences in Total Relative Frequencies Between Stage 1 and Stage 2 (Stage 2 – Stage 1)

The differences in criteria frequencies revealed especially in Figure 6.4 suggest some possible change from Stage 1 to Stage 2. At Stage 1 the most crucial thing is whether a document is topically related to the participants' information needs and what exactly a study is about. To support their topicality judgment, the

participants tended to use criteria such as journal, author, and publication date to exclude unwanted records. During the second stage, the participants' overall central concern was whether a document was useful for their particular project. Specific interests included whether the design established a good protocol, the findings added to the participants' own research, the study was linked to their projects, and the paper was of high quality. Other criteria such as "Newness" ($d = 5\%$) and "Data" ($d = 5\%$) were also used more often at Stage 2. Criteria with no change ($d = 0\%$) include "Understandability," "Readability," "Level," "Background Information," "Helpfulness," and "Length of Article." Note that most of the criteria that did not change from Stage 1 to Stage 2 were also infrequently mentioned at both Stages.

Macro Level Analysis

Reconceptualized Categorization

While the micro level analysis provides insightful results for the use pattern of criteria in document selection decisions, it may be useful to also have a more general picture of the use of dimensions of criteria. The original conceptual framework, as outlined in Chapter 3, proposed a three-class structure, featuring *Topicality*, *Quality of Information*, and *Cognitive State*.

It appears on the basis of the microanalysis that there is a need to extend the original three-category classification to better describe the rich set of criteria employed by the participants in their document evaluations. Upon close inspection of the complete list of criteria mentioned by the nine participants involved in this

study along with the above analysis of their use, an eight-category scheme was developed. Such a scheme was created not only by investigating the actual context in which relevance criteria were employed, but also by studying participants' comments regarding the nature of their own research work and their needs for information. This new structure contains the original three classes, *Topicality*, *Quality of Information*, and *Cognitive State*, and five new categories: *Research Structure*, *Source Value*, *Affective Aspect*, *Utility*, and *My Study*. Table 6.7 below delineates this reconceptualized and extended grouping structure³.

³ In the text, tables or figures of this dissertation, the term "criteria class(es)" is used interchangeably as "criteria category(ies)."

Table 6.7
Reconceptualized Categorization of Criteria Used

Category Number	Category Name	Criterion Items
1	Topicality	Domain Is About Interest of the Study Topical Focus Topical Relatedness
2	Research Structure	Argument Concepts Conclusion Data Design Evidence of Effect Implication Interpretation Method Nature of the Study Population Practicality Procedure Research Assumption Results Sample Size State of Research Statistic Analysis Techniques Theoretical Model Treatment Variables and Constructs
3	Quality of Information	Accuracy Background information Clarity and Well-Written Completeness Didn't read Importance Insightfulness Level Quality and Value Readability Repeat Scope Sophistication Specificity Starting Point Strangeness

Table 6.7
Reconceptualized Categorization of Criteria Used (Cont.)

Category Number	Category Name	Criterion Item
3	Quality of Information	Suitable for Meta-analysis Title Indicativeness Trustworthiness Uniqueness
4	Source Value	Article Type Author Author Bias Cited Author Cited Frequently Cited in Preliminary Paper Classic Study Geographic Location Item Mentioned by Coauthor Journal Language Length of Article Only Title Available Publication Date Reference Referenced in Items Selected See Items Cited This
5	Cognitive State	Add My Knowledge Certainty Expectation Familiarity Informativeness Inspirational Interestingness Read Before Remembering Understandability Agreeability Newness Originality Support My View
6	Affective Aspect	Affective
7	Utility	Helpfulness Usefulness

Table 6.7
Reconceptualized Categorization of Criteria Used (Cont.)

Category Number	Category Name	Criterion Item
8	My Study	Influenced My Study Is What I Want Justification of My Study Link to My Study Personal Interest Similar to What I Do Would Cite Reading

The original three categories *Topicality*, *Quality of Information*, and *Cognitive State* are numbered as Category 1, Category 3 and Category 5 in the table above. There are five criteria under *Topicality*, “Domain,” “Is About,” “Interest of the Study,” “Topical Focus,” and “Topical Relatedness.” *Quality of Information* embraces 20 elements, among them, “Accuracy,” “Clarity and Well-Written,” “Completeness,” “Level,” “Quality and Value,” “Readability,” “Scope,” “Sophistication,” “Trustworthiness,” “Uniqueness.” Fourteen criteria are grouped into the *Cognitive State* category. These criteria all have to do with users’ knowledge state. Criteria such as “Add My Knowledge,” “Familiarity,” “Informativeness,” “Inspirational,” “Interestingness,” “Understandability,” “Agreeability,” “Newness,” and “Support My View” are in this category.

This group of participants applied multiple criteria pertaining to the research method of a study. These criteria were so specific and detailed that it did not seem useful to simply group them under *Topicality*. Therefore, these criteria were separated into a new class of criteria called *Research Structure*. This new class covers

a range of criteria that relate to the various aspects of the research mechanisms of a study. This category is a rich set, containing a total of 22 criteria. Typical criteria are “Data,” “Design,” “Evidence of Effect,” “Method,” “Population,” “Procedure,” “Results,” “Sample Size,” “Statistical Analysis,” “Techniques,” “Nature of the Study,” and “Variables and Constructs.”

The original model did not include criteria that are related to the source of the documents. The nine participants frequently used such criteria as “Author” and “Journal.” Consequently, *Source Value* was established as an independent category. The criteria within this category all have to do with the source of the information items. For example, “Author,” “Journal,” “Language,” “Article Type,” and “Length of Article,” all represent characteristics of the source.

Criteria that describe users’ emotional reaction to the documents or their pragmatic value judgments of the documents are not included the original model. The participants’ verbal comments consisted of some affective reactions and utility evaluations. As a result, two new classes emerged: *Affective Aspect* and *Utility*. These two classes include the criteria “Affective,” and “Usefulness” and “Helpfulness” respectively.

A category that may be particular to doctoral students in psychology is the category that I label as “My Study.” *My Study* is a set of criteria that connects the relevance of documents at hand with the studies/experiments that the participants were performing. Seven out of nine people were conducting experiments on their topics, and, therefore, there were quite many criteria coming from these participants

that stress the connections between the documents reviewed and the participants' own studies. For example, "Justification of My Study," "Influenced My Study," "Link to My Study," and "Similar to What I Do." These criteria seemed to have strong properties of their own and point to a separate class. The *My Study* class provides evidence of the way that these participants shaped what was relevant in terms of their own research needs.

My Study is similar to the relevance facet *Oneself* in Cool, Belkin, and Kanter (1993)'s work. Cool et al. define *Oneself* as the "relationship between person's situation and the other facets" (p.79). Here, *My Study* has a strong emphasis on salience of the document read to the research projects that the participants were conducting. The criteria under *My Study* differ from the elements in *Topicality* in that these criteria do not describe the topical content of the documents, rather, they describe the aspects of the study reported in a text in relation to the participants' own research. The relationship may be topical, or it may not be topical. In fact in many cases, it was the design aspects or methodological issues that made the participants mentioning "My Study." However, these criteria were not grouped under *Research Structure* either, simply because the focus is mainly on *My Study* or the connections to *My Study*.

Data Processing

At this point, the original frequency counts based on the 15 criteria were no longer appropriate, due to the duplicate counts of criteria within one class. For instance, for a given document, a participant could have declared that it is both in

the right domain and is related to the topic. In this case the item would be coded both by “Domain” and by “Topical Relatedness.” Thus, when counting the frequency of criteria on a macro, categorical level for this example, the item should only be considered as using the broader criteria class *Topicality* once instead of twice. This required that the data be reprocessed by using node addresses to identify and eliminate duplicated codes. The relative frequency counts for the use of criteria classes were generated for each participant and for both Stage 1 and Stage 2 after the removal of duplicates.

Use of Criteria Classes by Participants

Table 6.8 lists the total relative frequencies of Stage 1 for all eight categories by the nine participants. Table 6.9 lists the average total relative frequency rates of Stage 2 by the criteria classes and by the participants. These two tables reveal use patterns for each criteria class by each participant. For example, at Stage 1, Participant 1 applied *Cognitive State* 55% of the time, while at Stage 2 the frequency increased to 67%. Participant 2 employed *Topicality* 27% of the time and *Source Value* 48% of the time during the record evaluation. And during the full-text evaluation, she used *Topicality* 10% of the time and *Source Value* 16% of the time. The third example is Participant 7, who used *My Study* only about 4% of the time at Stage 1. At Stage 2, she used the class much more frequently, at about 53% of the time.

Table 6.8
Use of Criteria Classes by Participants at Stage 1

	<i>Topicality</i>	<i>Research Structure</i>	<i>Quality of Information</i>	<i>Source Value</i>	<i>Cognitive State</i>	<i>Affect Aspect</i>	<i>Utility</i>	<i>My Study</i>
Participant 1	56%	7%	10%	41%	55%	0%	1%	1%
Participant 2	27%	68%	68%	48%	63%	8%	17%	25%
Participant 3	57%	72%	2%	11%	4%	0%	0%	0%
Participant 4	91%	3%	25%	16%	5%	0%	0%	5%
Participant 5	63%	47%	23%	30%	18%	0%	2%	4%
Participant 6	72%	35%	11%	17%	9%	0%	0%	9%
Participant 7	77%	13%	13%	28%	5%	0%	0%	4%
Participant 8	57%	39%	8%	51%	41%	0%	18%	14%
Participant 9	57%	20%	11%	23%	15%	10%	1%	15%

Table 6.9
Use of Criteria Classes by Participants at Stage 2

	<i>Topicality</i>	<i>Research Structure</i>	<i>Quality of Information</i>	<i>Source Value</i>	<i>Cognitive State</i>	<i>Affect Aspect</i>	<i>Utility</i>	<i>My Study</i>
Participant 1	87%	60%	50%	43%	67%	17%	57%	14%
Participant 2	10%	82%	48%	16%	30%	2%	24%	20%
Participant 3	57%	65%	22%	22%	28%	0%	5%	17%
Participant 4	75%	34%	7%	18%	3%	0%	0%	15%
Participant 5	51%	85%	28%	19%	23%	0%	11%	38%
Participant 6	100%	100%	36%	14%	14%	0%	50%	0%
Participant 7	8%	79%	36%	15%	50%	0%	28%	53%
Participant 8	48%	83%	9%	11%	26%	0%	37%	20%
Participant 9	63%	37%	10%	24%	10%	3%	17%	14%

Use of Criteria Classes by Stages. Figure 6.5 illustrates the information contained in Table 6.8 — the use of criteria categories at Stage 1 by the nine participants and by the eight criteria classes. Similarly, Figure 6.6 provides a graphical display of the information contained in Table 6.9 — the use of criteria classes at Stage 2 by the participants and by the categories.

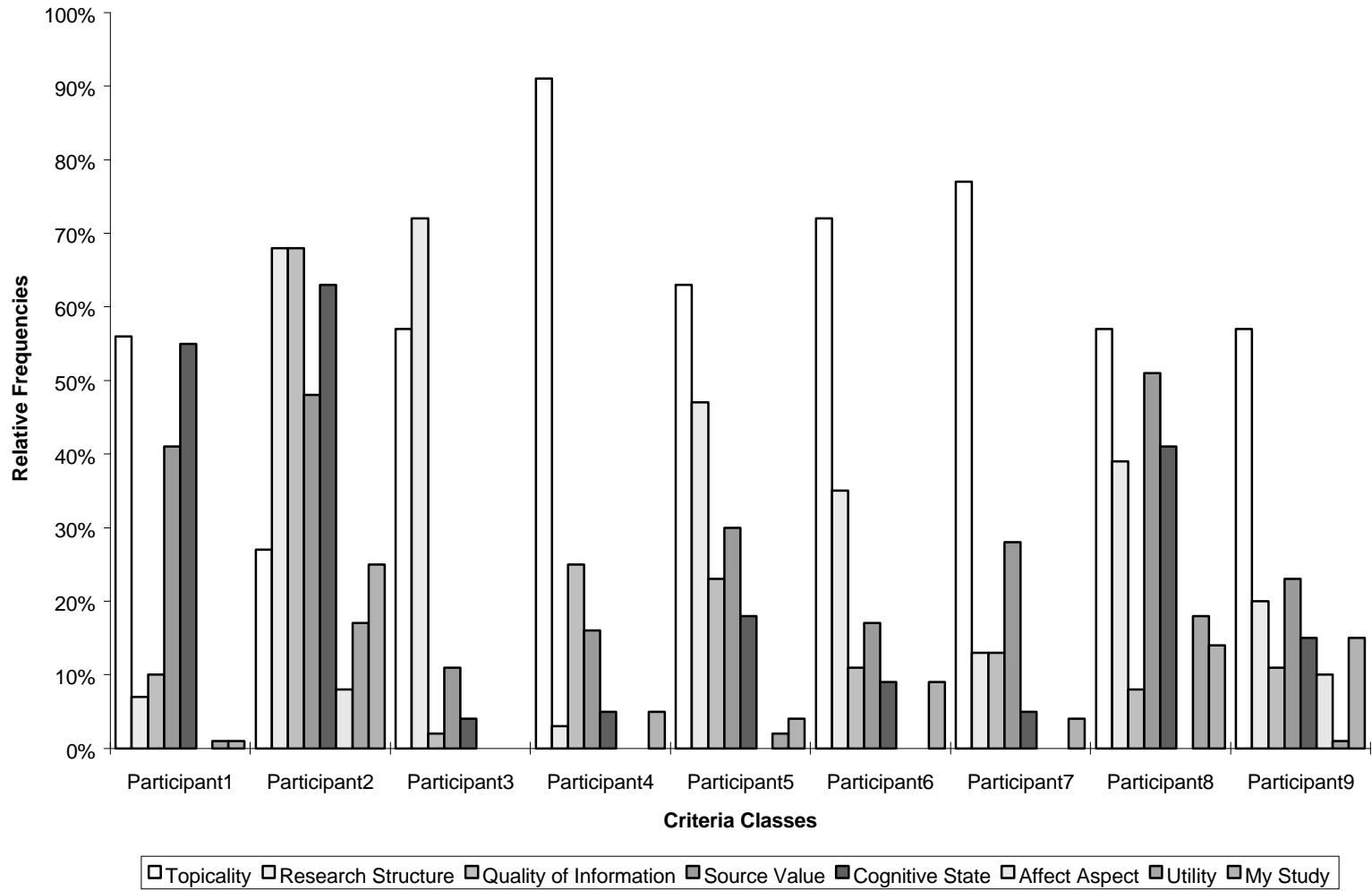


Figure 6.5. Use of Criteria Classes Stage 1 by Participants

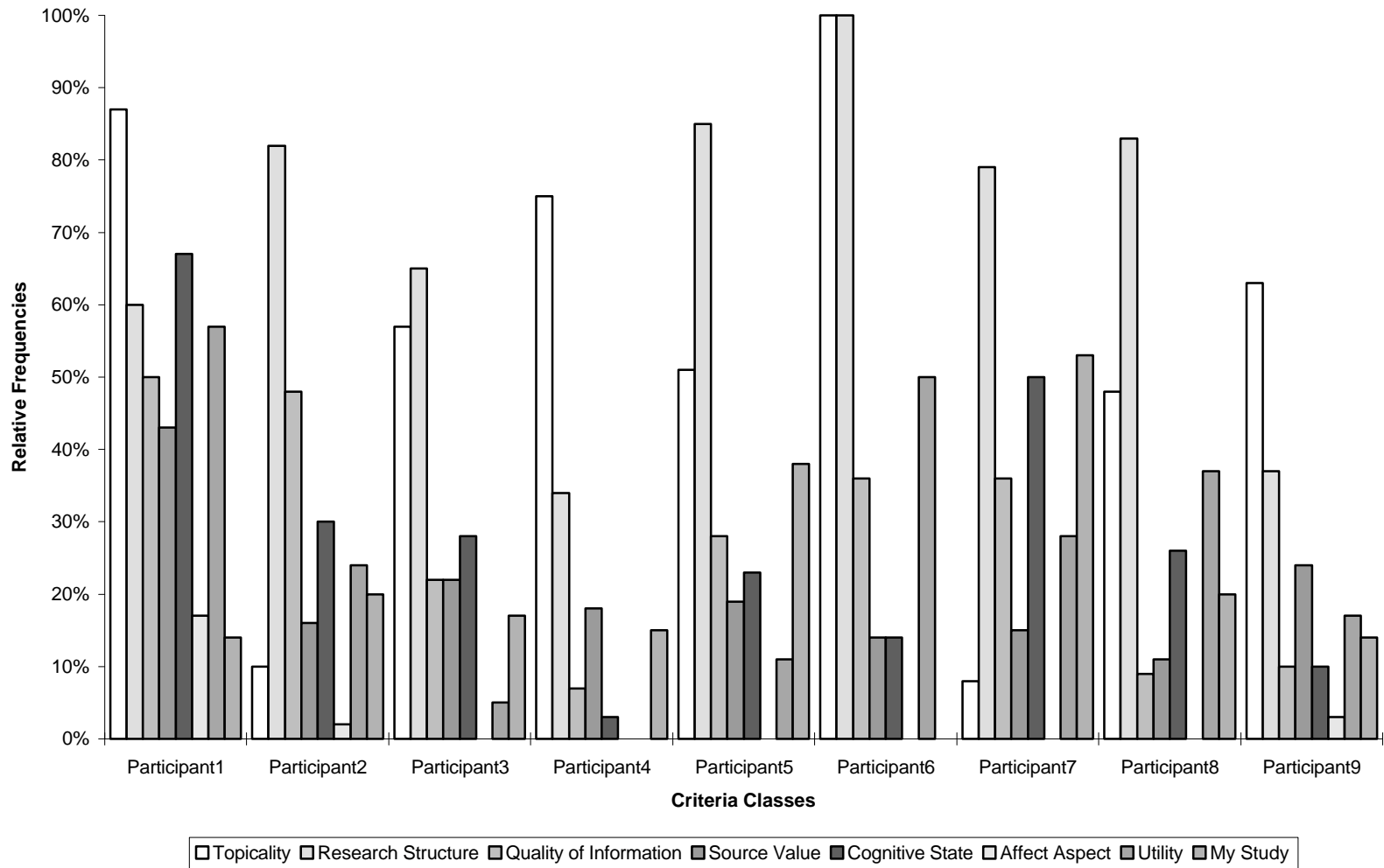


Figure 6.6. Use of Criteria Classes Stage 2 by Participants

From both Figure 6.5 and Figure 6.6 it appears that within each stage the use of criteria class is different from participant to participant. Roughly speaking, at Stage 1, *Topicality* had the highest frequency values for seven participants, except for Participants 2 and 3. For both of these cases, *Research Structure* was used most frequently. While *Research Structure* was used frequently for about half of the participants, Participants 1 and 4 use this set of criteria much less frequently. Participant 4 presented a sharp contrast between his use of *Topicality* (freq. = 91%), which reaches the highest percentage among all participants, and his use of *Research Structure* (freq. = 3%), which is the lowest frequency ratio among the nine people. During the post evaluation interview, Participant 4 claimed that “Domain” is the most important element in his decision making. This may explain why there was a strong emphasis on *Topicality* for both stages. On the other hand, this participant evidently considered little about the research method at the point of record evaluation. This may be at least partially due to the limited information that he could obtain from the records (recall that he encountered many records without abstracts).

For all participants, *Source Value* has a medium to high frequency rate. Participants 1, 2, and 8 applied *Source Value* more frequently than the others. These three participants also used *Cognitive State* frequently, but the class as a whole has medium to low frequencies for the others. Both *Quality of Information* and *My Study* have medium to low frequencies across all nine participants.

At Stage 2, *Topicality* remains the most highly used class for most of the participants except for Participants 2 and 7. These two participants concentrated more on the *Research Structure* of the documents. *Research Structure* became paramount for most of the participants, except for 4 and 9, who used *Research Structure* with a medium level of frequency. Even in these cases, while *Research Structure* was not used often during Stage 1, the frequency rates increased for Stage 2.

Both *Quality of Information* and *Source Value* hold medium to high frequencies for most of the participants. *Cognitive State* was medium to high for most people, except for Participant 4. Participant 4 did not use any of the *Cognitive State* criteria very much during record evaluation either. *Utility* was used at a medium level by most of the participants, so was *My Study*. Both classes generally have increased frequencies at this Stage. Lastly, *Affective Aspect* was used infrequently for both Stage 1 and Stage 2. At Stage 1, only Participants 9 and 2 used that class; at Stage 2, Participant 1 joined Participants 2 and 9 in using the *Affective Aspect* class.

Change in the Use of Criteria Classes by Participants. For Participant 1, all eight classes had increased frequencies at Stage 2. While *Topicality*, *Source Value*, and *Cognitive State* were used most frequently across the two stages, *Utility* ($d = 56\%$), *Research Structure* ($d = 53\%$), and *Quality of Information* ($d = 40\%$) increased greatly from low frequencies at Stage 1 to high frequencies at Stage 2. The least increase was for the *Source Value* class ($d = 2\%$). Recall that one of the goals for Participant 1 was to seek new design ideas, and as a result he emphasized strongly

the *Utility* and *Research Structure* of the studies. Both classes were used over 50% more frequently at Stage 2. *Topicality* also increased by 31%; this manifests the participant's intention of building a better knowledge base on the topic.

Quite opposite to Participant 1, Participant 2 had decreased frequencies for six classes, and increased rates for the other two. The increased classes were *Research Structure* ($d = 14\%$) and *Utility* ($d = 7\%$). The largely decreased classes included *Cognitive State* ($d = -33\%$), *Source Value* ($d = -32\%$), *Quality of Information* ($d = -20\%$), and *Topicality* ($d = -17\%$). It is interesting that Participant 2 specifically mentioned the importance of *Cognitive State* and *Quality of Information* during the post evaluation interview, but in her actual use, these two classes had large reductions from Stage 1 to Stage 2 (as measured by frequency of mentioning). Recall that Participant 2 was very succinct and comprehensive in her record evaluation. In comparison, her full-text evaluation appeared to have reduced frequencies in most of the dimensions, including *Cognitive State* and *Quality of Information*.

Participant 3 used most of the eight classes more frequently at Stage 2. The only class that has a minor decrease at Stage 2 is *Research Structure* ($d = -7\%$). The increased classes include *Cognitive State* ($d = 24\%$), *Quality of Information* ($d = 20\%$), and *My Study* ($d = 17\%$). Participant 3 did not think that there were changes in the importance of criteria as she moved from Stage 1 to 2. In her actual use, the frequency of *Topicality* remained the same at Stage 2, serving partially as a support for her own perception.

Participant 4 had a relatively large increase in his use of *Research Structure* ($d = 31\%$) at Stage 2. As he pointed out during the post evaluation interview, he employed “Data” as a new criterion for Stage 2 evaluation. He believed that whether or not a document contained data would determine whether he would be able to use the document for meta-analysis purpose. This may also explain why at Stage 2, Participant 4 put an increased emphasis on *Research Structure* and *My Study* ($d = 10\%$). As reported previously in the microanalysis section, Participant 4 used the criterion “Link to My Study” more frequently at Stage 2, causing the increase in the use of the class *My Study*. On the other hand, Participant 4 used *Topicality* and *Quality of Information* less frequently at Stage 2.

In Participant 5’s case, *Topicality* ($d = -12\%$) and *Source Value* ($d = -11\%$) had reduced frequencies at Stage 2, while *Research Structure* ($d = 38\%$) and *My Study* ($d = 34\%$) increased substantially from Stage 1. *Utility*, *Quality of Information*, and *Cognitive State* all had relatively small increases at Stage 2. Participant 5 believed that *Topicality* was the most important criteria for both record evaluation and full-text evaluation, but his actual use indicates a decrease in *Topicality*, with *Research Structure* rising to the most frequently used class at Stage 2. This participant also applied *Source Value* less frequently at Stage 2. On the micro level, Participant 5 used research related criteria such as “Design” and “Similar to What I Do” more often at Stage 2, and that is consistent with the macro level findings: both *Research Structure* and *My Study* increased greatly at Stage 2.

Participant 6 only reviewed seven full-text articles. Recall that he used very different sets of criteria for Stage 1 and Stage 2 evaluations and that he only had four common criteria across the two stages. On the macro level, two classes were less frequently used at Stage 2: *My Study* ($d = -9\%$) and *Source Value* ($d = -3\%$). The remaining classes all had higher frequencies at Stage 2. Among these were *Topicality* ($d = 65\%$), *Utility* ($d = 50\%$), *Research Structure* ($d = 28\%$), and *Quality of Information* ($d = 25\%$). At Stage 2, Participant 6 paid greater attention to criteria such as “Variables and Constructs” and “Statistical Analysis,” and this may explain why the use of the class *Research Structure* was largely increased. For Stage 1, Participant 6 generally looked at the “Topical Relatedness” for most of the records, however, at Stage 2, each of the seven articles were examined by the criterion “Topical Focus.” The relatively frequency for *Topicality* was thus increased.

The degree of change in the use of criteria classes was relatively large for Participant 7, going from as high as 69% to as low as 13%. However, during her post evaluation interview, the participant stated that she did not feel that there were changes in the use or importance of the criteria. Her actual use pattern seemed to contradict her after-the-fact reflections. All seven classes that she used (she did not use the class *Affective Aspect* at both stages) changed greatly from Stage 1 to Stage 2. The highly increased criteria classes included *Research Structure* ($d = 66\%$), *My Study* ($d = 49\%$), *Cognitive State* ($d = 45\%$), *Utility* ($d = 28\%$), and *Quality of Information* ($d = 23\%$). The decreased criteria classes were *Topicality* ($d = -69\%$) and *Source Value* ($d = -13\%$).

Building on the micro level results, it seemed that Participant 7 was looking for documents that contained “Link to My Study,” or applied a good “Treatment” and “Technique,” or include conceptual and methodological “Newness,” and that may be why use of most of the related classes increased at Stage 2. The decrease in *Topicality* may suggest that at Stage 1 the participant’s decision-making was focused on *Topicality*, i.e., whether the document described her topic “cognitive therapy and OCD.” However, at Stage 2, granted that the selected articles fulfilled the topical requirement, *Topicality* was no longer a crucial element when compared with other important factors such as *Research Structure*, *My Study*, and *Cognitive State*. The increase in *My Study* may also be contributable to the fact that at Stage 2, the participant was purposefully rating usefulness based on how much the study reported in a text was connected to her own research and whether that study influenced her design.

For Participant 8, the category with the greatest decrease in use from Stage 1 to Stage 2 was *Source Value* ($d = -40\%$). Use of both *Cognitive State* ($d = -15\%$) and *Topicality* ($d = -9\%$) also decreased at Stage 2. *Research Structure* increased by 44%, due to the fact that Participant 8 used “Design” and “Nature of the Study” more frequently at Stage 2. *Utility* ($d = 19\%$) and *My Study* ($d = 6\%$) were also used more often at Stage 2. Evidently, at Stage 2, the participant concentrated on how the study reported in a text would help her to develop her own research project.

Participant 9 demonstrated a relatively stable use of criteria classes. *My Study*, *Source Value*, and *Quality of Information* were used with almost equal

frequency at both stages. *Research Structure* increased by 17% while *Utility* increased by 16%. *Topicality* ($d = 6\%$) also increased at Stage 2. On the other hand, both *Affective Aspect* ($d = -7\%$) and *Cognitive State* ($d = -5\%$) were used less often. As reported earlier, Participant 9 focused on “Usefulness,” “Nature of the Study,” “Theory,” and “Author” at Stage 2. This is consistent with the findings that *Research Structure* and *Utility* were increasingly used at Stage 2. Additionally, the fact that *Source Value* did not decrease at Stage 2 may be related to the third criterion that the participant described himself using for the full-text evaluation. According to him, the third criterion he used was that a key author wrote the documents. Consequently, “Author” was used more often at Stage 2 and the related class *Source Value* ($d = 1\%$) had a small increase at Stage 2.

As for the micro level description of the use of individual criteria, the macro level examination of the use of criteria classes by each participant suggests that the use is very situational. However, the participants’ general use pattern often agrees with their use of individual criteria, which in turn, relates to the participants’ search interests and their specific needs for information.

Use of Criteria Classes Across Participants

Calculations were made to summarize the total relative frequency for each criteria class across participants. For a given class, the total relative frequency was obtained by adding the raw frequency of all nine participants, and then normalizing it by the total number of documents all nine participants had reviewed. For Stage 2, the average of total relative frequencies of written comments and oral comments

was used to represent the total relative frequency for that stage. Table 6.10 lists the total relative frequencies for Stage 1 and Stage 2, and Figure 6.7 illustrates the corresponding data structure contained in Table 6.10.

Table 6.10
Use of Criteria Classes across Participants Stage 1 and Stage 2

Criteria Category	Total Relative Frequency Stage 1 (N=753)	Average Total Relative Frequency Stage 2 (N=246)
Topicality	63%	49%
Research Structure	32%	67%
Quality of Information	20%	25%
Source Value	28%	20%
Cognitive State	24%	27%
Affective Aspect	2%	2%
Utility	3%	21%
My Study	8%	25%

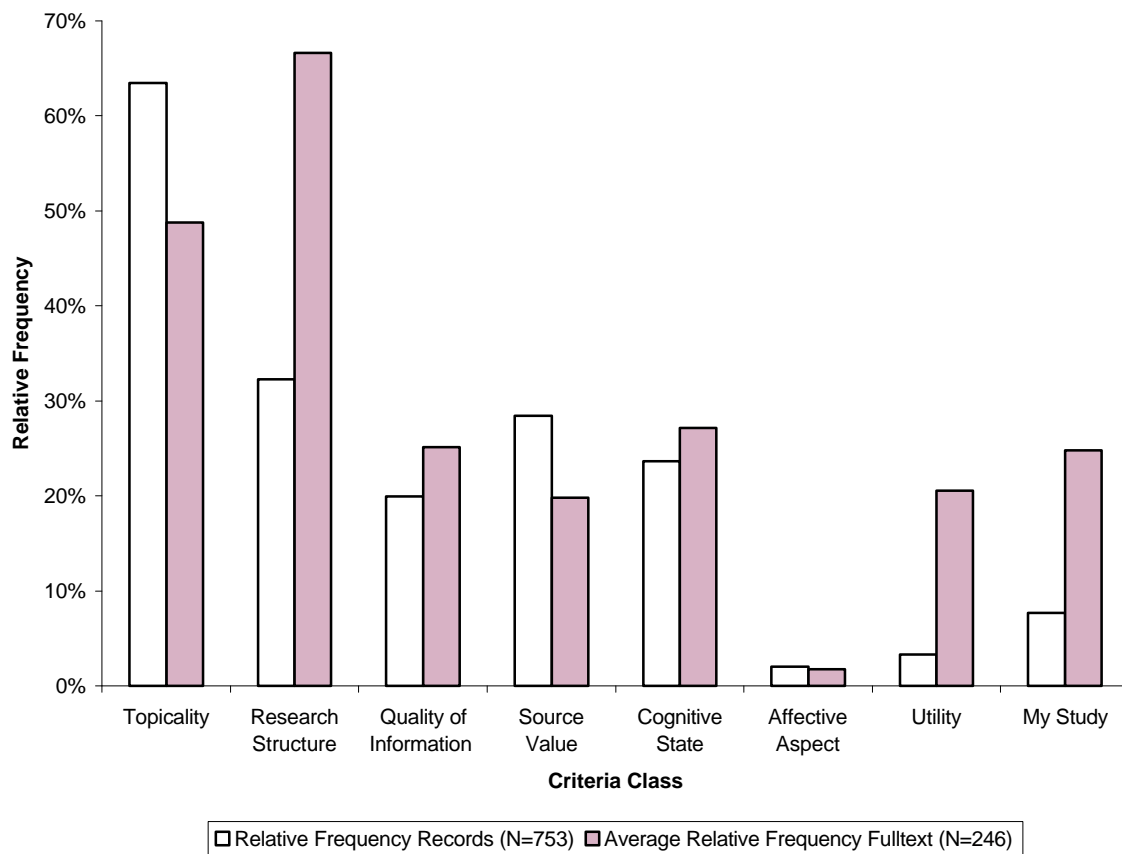


Figure 6.7. Use of Criteria Classes across Participants Stage 1 and Stage 2

Across participants, the most frequently used criteria class at Stage 1 is *Topicality*, with a total relative frequency of 63%. *Research Structure* soared up as the most frequently used criteria class at Stage 2, with a rate of 67%. *Topicality* was still the second most highly used criteria category at Stage 2, with a relative frequency of 49%. *Quality of Information* (freq. Stage 1 = 20%, freq. Stage 2 = 25%), *Source Value* (freq. Stage 1 = 28%, freq. Stage 2 = 20%), *Cognitive State* (freq. Stage 1 = 24%, freq. Stage 2 = 27%) had a medium frequency of use for both stages. *Affective Aspect*, on the other hand, was the least frequently used category for both stages — 2% of the time. *Utility* and *My Study*, were also relatively low in their frequency values at

Stage 1, however, at Stage 2 they both increased to usage rates of 20% and 25% respectively. The difference between the highest and the lowest frequency values is 60% for Stage 1 and 65% for Stage 2.

If using 25% (i.e., one quarter of the time) as the cut-off point for highly frequently used categories, the interpretation of the results becomes much more direct and simplified. At Stage 1, the participants relied on three major factors for their reasoning in evaluations. They primarily employed the criteria pertaining to the *Topicality* of the documents. In addition, they used the criteria that describe the *Research Structure* of the studies as well as documents' *Source Value* to determine whether to accept or reject a document. At Stage 2, the participants included more components to their decision making. This time, they depended greatly on *Research Structure*, while still keeping some aspects of the *Topicality* dimension in mind. Elements that were relevant to the needs of their *Cognitive State* also carried some weight in their evaluations, so did characteristics reflecting *Quality of Information* and connections to *My Study*.

Another perspective that describes the use of criteria classes across participants is obtained by ordering the eight classes in rank order according to their frequency values. Rankings of criteria classes were thus generated for Stages 1 and 2. Table 6.11 provides the rankings. The marked (*) criteria are those that are above the 25% cut-off point.

Table 6.11
Rankings of Criteria Classes Based on the Total Relative Frequency Rates

Criteria Classes	Stage 1	Stage 2
Topicality	1*	2*
Research Structure	2*	1*
Quality of Information	5	4*
Source Value	3*	7
Cognitive State	4	3*
Affective Aspect	8	8
Utility	7	6
My Study	6	4*

Note. * Criteria whose frequencies were above the 25% cut-off point.

At Stage 1, the most frequently used criteria classes were *Topicality*, *Research Structure*, and *Source Value*. At Stage 2, the most used became *Research Structure*, *Topicality*, *Cognitive State*, *Quality of Information*, and *My Study*. *Source Value* dropped to seventh at Stage 2, ranking the second to the lowest. The lowest ranked class for both Stages is *Affective Aspect*. Apparently participants did not use the *Affective Aspect* much for either stage. The second to the lowest ranked class for Stage 1 is *Utility*, which ranked sixth at Stage 2. While *Utility* and *My Study* were the second and third least used classes for Stage 1, *Source Value* and *Utility* were the second and third least used categories for Stage 2.

Change in the Use of Criteria Classes between Two Stages

The change in the use of criteria classes between the two stages was examined through two viewpoints. The first view is of the differences in relative frequency values of criteria classes between the stages. The second is of the differences in ranking levels of criteria classes between the stages.

Figure 6.8 illustrates the differences in relative frequency values between the stages. A positive bar indicates the frequency value at Stage 2 is higher than that of Stage 1, whereas a negative bar indicates the opposite.

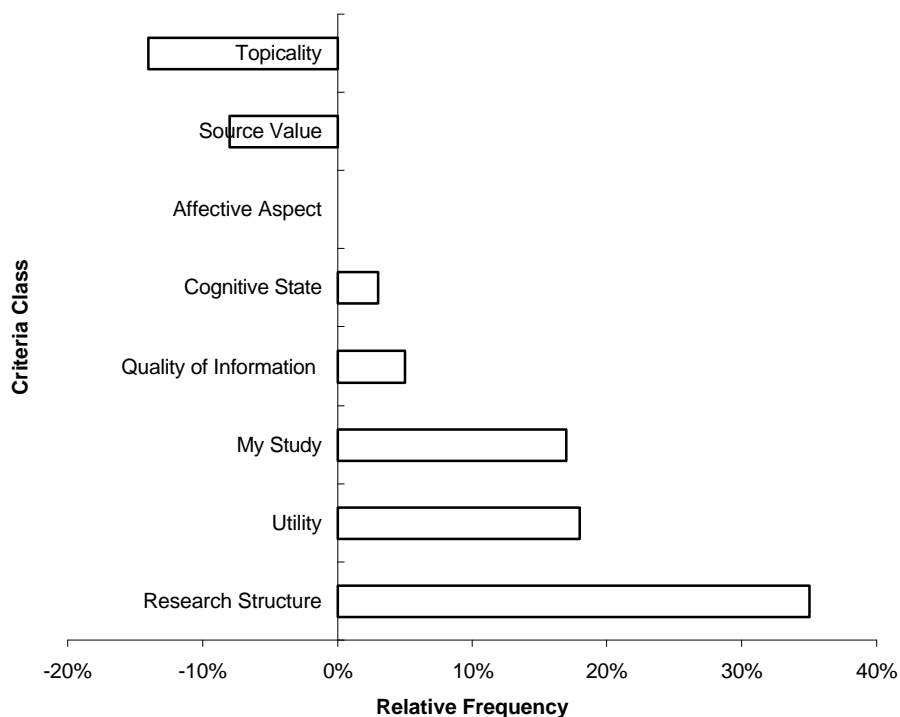


Figure 6.8. Differences in Total Relative Frequency between Stage 1 and Stage 2

It is apparent from the above figure that *Research Structure* had the greatest relative change among all eight criteria classes. Its frequency increased 35% at Stage 2. *Utility* and *My Study* also increased a good deal at Stage 2, with a positive rate of 18% and 17% respectively. *Topicality* ($d = -14\%$) and *Source Value* ($d = -8\%$) were the only two classes that were used less frequently at Stage 2. Both *Quality of*

Information ($d = 7\%$) and *Cognitive State* ($d = 5\%$) were applied a little more often at Stage 2. *Affective Aspect* had no change across the two stages.

All of this suggests that participants had much stronger interests in the research mechanisms of the documents at Stage 2. They also focused more on usefulness of the documents and the relationship between the studies reported and their own projects. Additionally, their concerns of the quality of the papers and fulfillment of their own cognitive needs were a little stronger than what they had at Stage 1. On the other hand, at Stage 2, the participants appeared to focus less on the topical and source aspects of the documents.

Figure 6.9 describes the change in the use of criteria classes through the view point of change in the ranked positions of the classes from Stage 1 to 2.

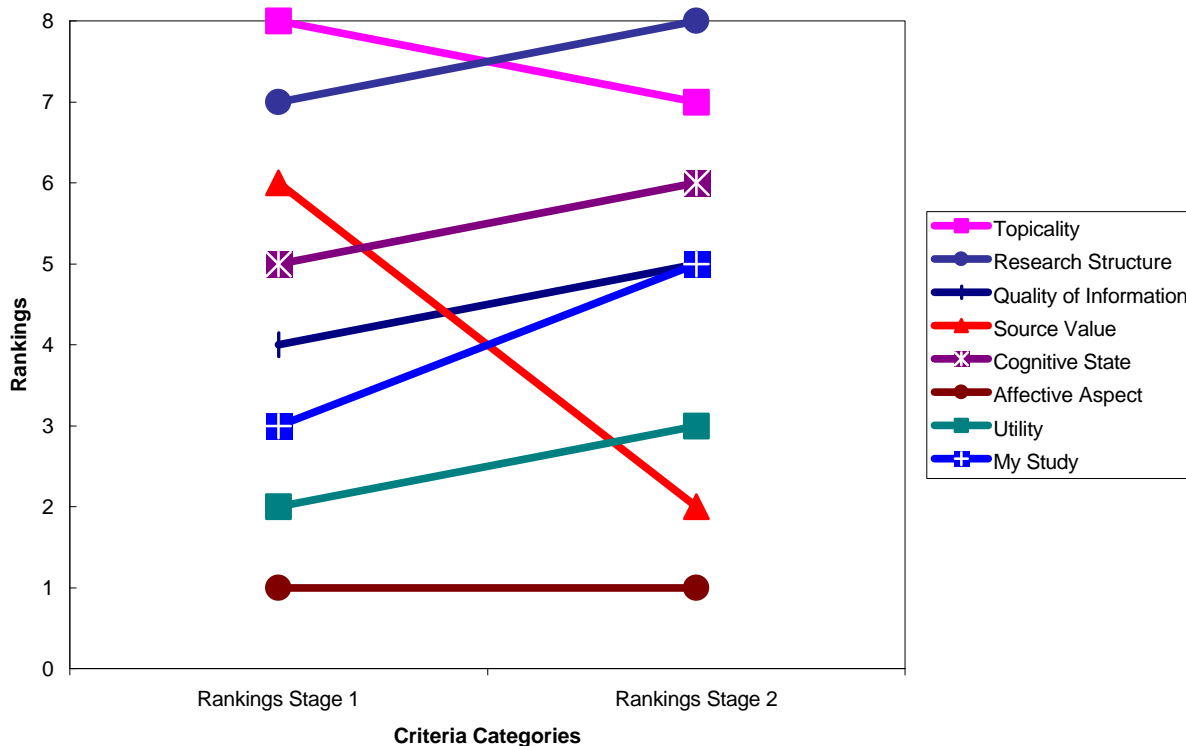


Figure 6.9. Differences in Rankings of Criteria Categories from Stage 1 to Stage 2

Note that at Stage 2, *Quality of Information* and *My Study* have the same ranking (fourth). The criteria class that projected the greatest change in the rankings is *Source Value*. It was ranked third at Stage 1 and dropped to seventh at Stage 2 ($d_r = -4$). The ranking position of *My Study* increased from sixth for Stage 1 to fourth for Stage 2. *Research Structure*, *Cognitive State*, *Quality of Information*, and *Utility*, all increased by one unit at Stage 2, whereas *Topicality* decreased by one unit from Stage 1. The only class that held a consistent ranking from Stage 1 to 2 was *Affective Aspect*, which ranked the lowest at both stages.

As measured by change in ranked order, *Source Value* had the highest change, dropping four units at Stage 2. The reduction in participants' appreciation of *Source Value* during full-text evaluation may suggest that participants used the criteria at Stage 1 as a threshold for including or excluding records and at Stage 2, most of the documents already passed the threshold so the participants no longer needed to address that aspect for their full-text evaluations. The two-unit increase in the rankings of *My Study*, on the other hand, may suggest that as participants examined full-texts, they consciously made the analog between the study reported and their own research a stronger indication of the value of the document reviewed.

Two general trends of change or transformation emerged as the participants progressed from Stage 1 to Stage 2. The first shift is the emphasis on *Research Structure* over *Topicality* at Stage 2. The second change is that *Source Value* was important at Stage 1, but at Stage 2 the focus shifted to *Utility*, *My Study*, *Quality of Information*, and *Cognitive State*. These two shifts indicate a pattern of change in reasoning. That is, a move from a somewhat objective and impersonal orientation to a more personal, situational, and subjective way of thinking.

The following section provides an account of the participants' comments on their use of criteria and the shifts and changes in the use of criteria as they advanced from one stage of document evaluation to another.

Participants' Perceptions of Criteria Use

Participants voiced their thoughts as they evaluated bibliographic records. During the full-text evaluation interviews, which used the participants' written comments as a point of departure, participants orally reviewed each article they read. The entire process (covering both Stage 1 and 2) contained many discussions about the heuristics that the participants employed during their evaluations and their views on the use of specific criteria. Further, during post evaluation interviews, participants were asked to think explicitly about their use of criteria for both stages. In the section below, the participants' comments are analyzed and arranged in relation to the specific criteria. For a comprehensive list of the definitions and examples of all the criteria used by the participants, refer to Appendix I.

Individual Criteria

Topicality. With little exception, participants all believed that "Topicality" or "Topical Focus" was the single most important criterion for their evaluations of bibliographic records. Using "Author" for comparison purposes, Participant 4 indicated that topical relevance is an essential, irreplaceable criterion for document selection. He said,

I wouldn't use "who is it written by" to make a final decision. That is just something that makes me look at things more closely before ruling it out. But if nothing else looks relevant, but the author was important in that area, I still wouldn't take it because of that. That wouldn't be a reason. It may be just I know to look at it closely before ruling it out. But a [author] name wouldn't really add anything if it wasn't relevant. The topic is the most important thing, yeah, definitely.

To participant 4, “the topic” is an overriding factor that determines the selection of the items, whereas “Author” only serves as a secondary criterion. It is used to enhance the selection judgment based on topicality. It could lead to a more in depth consideration of an abstract, but there would also need to be some topical connection in place initially. Other participants also shared the same view of the function of criteria such as “Author” or “Journal” (see those sections below).

Participant 5 also believed that “Topical Focus” is the most important criterion for his decision making at Stage 1. He pointed out “for the most important, I think it is the ‘Topical Focus.’ I mean just in terms of how often I used that, and how many articles that I would comfortably have excluded purely on the basis of ‘Topical Focus’.”

At Stage 2, “Topicality” was considered as equally important or even more so. Most people indicated that “Topicality” remains the top criterion for decision making. For example, Participant 1 suggested that when reading actual articles, “topicality is the most important, and it’s more important than it was.” Participants 3, 5, and 8 all declared “Topicality” to be the most important criterion for both Stage 1 and Stage 2. Participant 3 stated, “so whether I am looking at it on the abstract and title, or reading the actual article, the most important thing is that it involved the topic that I am interested in.” Participant 5 said, “I am trying to think what's the most important, I still think the single most important criterion at the document stage is ‘Topical Focus’.”

On the other hand, Participant 8 indicated that there was a slight difference in what “Topicality” meant at Stage 1 versus what it meant at Stage 2. She pointed out,

Yeah, that [topicality] was the most important one. Although there was sort of difference in what sort of topical things I was looking for at Stage 2. I was much more interested in justification than I had been in Stage 1.

Journal. During Stage 1, most of the participants used the “Journal” criterion to help them to make selection decisions. In evaluating one of the records, Participant 9 pointed out, “My rule of thumb is that anything in [a journal name] is garbage.” When asked why, he responded,

You can publish anything in there just as long as you pay for it. And in my opinion the quality of the research paper you have, is partially based on the article that you cite, I don't think it's something I want to do. I don't want cite [a journal name]'s article.

Participant 1 made the following comment after he reviewed a group of bibliographic records,

Actually, what I am using to decide: First, if the title looks extremely relevant, I will be interested; if it looks a little bit relevant, I might look at the things like authors, are they publishing in the social psychology journals versus, what was the last one, [the journal name]. That might be relevant, but not as essential as other ones. I don't think I am using these criteria as a personal organized thought; I don't think other people do either, but that's how I decide.

When commenting on the use of the criterion “Journal” during the post evaluation interview, Participant 1 further stated,

We should remember also that when I did the search, I sort of selected the articles that tended to be in good journals... And on

average, I think there is a reason why those things are in good journals. But I am not sure the extent to which I might be biased by the journal. So if I found something looks kind of good, but it is in [a journal name], I will take it; if it's in another journal, I wouldn't.

From above statements, it seems that Participant 1 used “Journal” to assist his otherwise not too certain decisions. Apparently, “Journal” was not used as the sole determinant for his document selection. In the actual decision making process, “Journal” is often accompanied with other criteria, especially some primary criteria such as “Topicality” and “Domain.”

Participant 3 also believed that “Journal” and “Author” are not as essential as some other criteria such as “Topicality.” She labeled “Journal” and “Author” as the “superficial” criteria, whereas the more important criteria would be “Topicality,” “Nature of the Study,” and “Measurement.” Below was her exact statement while evaluating a record:

Yeah, this is a yes. I mean superficial things are author, journal title, and the more important things are the reciprocal effect being there, the longitudinal nature of the study, and the really unique way of measuring stress and coping.

Participant 1 also observed that as he moved to Stage 2, “Journal” as well as “Author” became less important. Here is what he said about this issue:

Also journal was extremely important then [Stage 1], a little less important now [Stage 2]. Journals are a really good indicator, because I know the typical caliber of research that gets into certain journal, so caliber of research is important, but also relevance of research, I know [a journal name] is basic site of social psychology, whereas [a journal name] is an applied site of social psychology, and they are going to have, probably be asking different questions with perhaps different methodology.

So again, those two [“Journal” and “Author”] I think were more important at first.

A further interpretation of Participant 1’s above comment is that both “Journal” and “Author” may be used to help people to guess topical relevance and other aspects of the item. Both “Journal” and “Author” serve as good indicators of various things, including the topicality of the item, the nature of the study, and the quality of the document, and even the characteristics of the design at a gross level. At Stage 1, with the limited information in a bibliographic record, an educated decision can be made by employing things like “Journal” and “Author” to aid an otherwise incomplete decision. The situation for Stage 2 is different, and more discussion of this in later sections.

Author. Participants frequently made comments of the use of “Author” in combination with the use of “Journal.” Most of the participants seemed to consider these two as the same sort of criteria that held more functionalities at Stage 1 than they did at Stage 2. Participant 4 elaborated a little bit on how he used “Author” as a criterion for record evaluation.

I will take a little more time with the familiar authors than not with unfamiliar authors, right. In the case of ambiguity, if I wasn't sure whether I'll keep it, and it was a famous author, then I would make sure that I would really did not want to use that article. I would take a little more time.

At Stage 1, “Author” served as a criterion to help the participants to decide on a record whose topical relevance was not obviously shown. Participant 5 provided a similar explanation for why “Author” would be more useful at Stage 1.

He said, “it seemed again to be useful for like making as educated guess as I could have for what the article is going to be about, how well written it's going to be.”

This suggests that “Author” had a special role at Stage 1 in assisting participants to make inferences on at least two aspects of an item: the topicality of the document and the quality of the study. As evidence of this, the following is an excerpt of Participant 5’s comments as he evaluated a record:

I am almost certain this is relevant. It talks about Big-5, [a journal name], major article, [an author’s name]. Some of the authors are original zero order stuff...

At Stage 1, “Author” was also helpful when the participants were not familiar with the study described in a record. Participant 1 commented,

I think that when I didn't know very much about the research, like when I was doing the search originally, I had to say, pretty much, well, these are the names that I know are very big, what have they done. So authorship was huge early on, and is less important now [full-text evaluation].

Participant 5 also believed that at Stage 2, “Author” was not as important. “And once I have the full-text, I guess I didn't really need that kind of indirect information about how relevant it's going to be, or how useful they were all written. So yeah, author takes a lesser role.” With the full-text in front of them, the participants no longer felt the need of making guesses from other sources about the topicality or the quality of the article. Therefore, once an important function of the “Author” criterion at Stage 1, i.e., providing hints and helping people making inferences of the document, no longer exists for Stage 2. At Stage 2, the participants still mentioned “Author” in their evaluations, but for the purpose of supporting

their points during the evaluation. Below is an example of how Participant 5 used the criterion “Author” at Stage 2.

...and this article, one of the coauthors is really well-known statistician, and they did a really good job of two things, that is, one laying out very clearly what conceptually it means if a set of variables follow the circumplex.

Participant 4 self-observed more strongly about the change in importance of “Author.” In answering to my question during the post document evaluation interview, he provided a detailed description on why “Author” weighed less for the second stage.

So the first thing you were asking is after I read the article, did the criteria change? I'd say yeah, in some way they do. Like “Author” will become less important, because I am getting more information about the paper and if it's someone who's not very well-known, yet they are providing very compelling evidence in a way that's highly relevant to my topic area, I will certainly keep the article. And likewise if a famous author with a compelling title had a very weak paper, I would be inclined to not keep it after reading it.

Meanwhile, Participant 8 did not view “Author” to be an important criterion for any of the two stages. She commented, “‘Author,’ again, that didn't really matter once I had it all in front of me.... But in general that wasn't something I paid too much attention to.”

Publication date. Most of the participants realize the importance of the recency of a document; however, they also suggested that as for “Journal” and “Author,” “Recency” became less important at Stage 2.

As he was reading through a bibliographic record, Participant 5 was hesitant to select the item, since he could tell it is pretty old. Below is what he had commented,

If I didn't know there is other research done on first impressions, I may be attempted to take this, but it's fairly old. I will take it for now, but if I find other things that explicitly on first impressions, I will probably drop it. Just because it is so old, it doesn't make it so bad, it's just nothing like the most recent material.

Some participants used “Publication Date” as one search step to rule out the unwanted out-of-date items. Others, based on the publication year of articles, identified those that were the classics or the first studies on the topic. For instance, during the evaluation of a record, Participant 2 stated, “...so this would be a really, really good, actually probably an excellent starting point because it is a 1996 review so it is relatively recent.” During the evaluation of a full-text article, Participant 2 mentioned the “Publication date” to highlight the importance of an original study. Below is her comment as she described a full-text article.

This article was the bomb, but for very odd reasons. I rated it as a 7, it is extremely useful because the ideas of preemptive analgesia have been basically associated with a man named [a researcher's name] in 1980. This guy [the author's name] was at 1913, and he did it first. And I did not know that. So what I wound up with this article sent me back to his original stuff in 1913. He's written a book about it, it was great, and actually the thing that is really exciting about this is this guy wrote it in 1913 about a lot of these designs. And it was really strange, like pre-incision versus post-incision and why it would work as opposed to, and it's just brilliant stuff.

Depending on the topics, the importance of “Recency” varies by the participants. This criterion generally carried less significance at Stage 2. Participant

8 defines “Recency” as a “watching” criterion. Her exact words were, “In terms of actually reading the papers, ‘Recency’ kind of, it wasn't as important...this was my watching criterion.”

Newness and Inspirational. The criterion “Newness” differs from “Publication date” or “Recency” in that it describes whether a given item contains information (concepts, theories, design, or analytical technique, etc.) that is new or novel to the participants. In this study, “Newness” was grouped under the *Cognitive State* category, while “Publication Date” belongs to *Source Value*.

Typically, the participants’ evaluations would be coded as a negative value under the criterion “Newness” if the participants indicated that they already knew the issues covered in the documents, or the documents added nothing new to them. One unique thing about this group of participants is that not only did they use the criterion “Newness,” but also they highly praised this criterion during the discussions and interviews. In the post evaluation interview, one participant proposed that all the other criteria that she used could somehow be integrated under “Newness.” In her view, “Newness” serves as an overriding, catchall type of criterion.

One thing I am thinking, I guess, some of these things are important in that they increase the probability that the issue of newness, which I think is paramount, will be there. So for example, I know that certain authors do really good work, and so are very creative and I think I can learn from them, so if I scan for certain authors, it's more likely it will give me this newness that they will say something useful to me. So it doesn't end there, it's not like I want you know, it is not an end point. It's just sort of means to this end I have. That give me some information, ultimately this [the newness criterion] is the only thing I want. So it has to be relevant and new.

Participants considered that the importance of “Newness” increased greatly at Stage 2. Participant 1 believed that he hardly used “Newness” during record evaluation but it became salient to his Stage 2 evaluations. He said, “Theoretical newness, I really don’t think was much of a criterion at all at first, and I think, as I read, it became more important.” As he reflected on his evaluations of full-text articles, he indicated that in some cases “Newness” would even makeup some inadequacies in topical relevance for the documents.

And there were a couple of articles that I gave very high scores to that fit that. They are relevant, but not totally relevant, but their theoretical newness and excitingness was enough to rate them pretty high.

Along with the criterion “Newness,” Participant 1 also appreciated the documents that are conceptually inspiring, or in his own words, “hone my thinking.”

All these were relevant in the sense that, they are not directly addressing this issue, but they were relevant in the sense that they were so brilliant and all-encompassing that I can, it honed my thinking regarding the project at hand. Even though it wasn't necessarily directly relevant.

According to several participants, at Stage 2 an article would receive a relatively lower rating if it did not give them something new. To explain his rating for one article, Participant 9 said, “This is sort of an early review article by [an author name]. And it has a lot of stuff that I already looked at, so I gave it a 6 rather than a 7. Because it's an early review, and also because I already knew a lot of

research he cited in it.” In a later discussion, Participant 9 pointed out the role of “Newness” at Stage 2, “I think newness is all that came into play later.”

Participant 2 also used “Newness” as a standard for full-text evaluation. She rated one of the articles very low, because it contained much of the information that she knew already. She commented,

I gave it a 1 or 2, not because it's a bad article, but just cause I knew all that already. So it was a great article, it was very well written, it was excellent. But it was a complete review of things that I already knew and was already familiar with, so it wasn't insightful in that manner.

In reflecting on overall use of the criterion “Newness” at Stage 2, Participant 2 stated,

When I read the individual article, it would be like does this add more information, or is this just basically something that I've already seen...but what really happens was just a big planting up with the detail. Particular in regard to the design, and toward the end of it, relevant was really what was new, and what wasn't new.

Overall, “Newness” appears to be an important criterion, highly valued by the participants. This may be related to the particular academic status of the participants, since most of them were at the stage of either planning or actual composing their dissertation proposals. Their attention was naturally focused on searching for new ideas, new theoretical frameworks, or new design methods to help them to modify or improve their own research. This particular criterion was very important at Stage 2, since it could be rather difficult based on the limited information provided through bibliographic records to judge that the item would add something new to them. Interestingly, even though the across-participant

frequency of the “Newness” at Stage 2 (freq. = 6%) increased from Stage 1 (freq. = 1%), the increment is not as large as it would be based on the participants’ verbal comments.

A clarification needs to be made regarding the distinction between “Newness” and “Read Before” (or “Document Novelty” in some researchers’ definitions). As suggested by the participants in some comments quoted earlier, “Newness” is closely associated with document novelty – the participants would know the information contained in an article if they read the article before. However, “Newness” is different from “Document Novelty” in that a document can be completely new to a person, yet it is a review of the studies or a restatement of a theory that the person is already familiar with. In such a case, “Newness” does not equal to “Document Novelty.” It is more equivalent to Barry’s (1993) “Content Novelty.” In the same way, “Recency” can sometimes serve as an indicator for “Newness,” but an article that has an old date could very well contain some new information or add something new to a person’s knowledge base. Therefore, “Recency” may be an indicator of “Newness,” but not necessarily the reverse.

Writing style and information presentation. At Stage 2, the participants were able to judge the usefulness of the full-texts based on writing style. This criterion was especially mentioned by two of the participants during the full-text evaluation. Participant 2 indicated that “one of the things that was most interesting in actually reading the articles that became a criterion ... is how the things were written ... Writing style and how the ideas were presented became extremely important ...” Her examples were from articles that involved “animal research.”

Part of the animal research is really hard to read, it is kind of like; I don't know how to describe it. It's very jargonish, very unapproachable, and they use complete different terminology for reasons that are ill sensed than a lot of human literature, it reads more stilled and more stiffened kind of... So it makes those types of articles a lot more difficult to employ.

Participant 5 used a criterion that is broadly related to an author's writing style but specifically referred to how well defined the concepts and theories was. According to his explanation, it was "sort of like the combination of writing style and explicitness or thoroughness in defining their concepts." He continued, "that kept, I think, at least two articles from being more useful than they could have been... I didn't use that in so many cases... but it worked against those two articles, and there was another article that it worked in favor of this..." The following continues Participant 5's discussion of his use of this criterion:

I don't know if it is because they are [the name of a nationality], and just there are some language differences in their writing style, but they were using a lot of terminology that seem to be very loose kind of terminology and they didn't define things very well, they used terms that sort of roughly seem to I could understand, but it could've two or three things, what they were saying could've meant, I mean the implications of what they are saying would've been very different if they meant this thing versus this other thing...So it certainly was relevant, but it was pretty frustrating.

Evidently the criterion "Writing Style and Presentation of Information" held some impact in the participants' evaluation of full-texts, but not for record evaluation. Authors' writing style can hardly be very well reflected through an abstract, for example.

Influenced my study. Another unique criterion that was mentioned by participants only during the second stage of the evaluation was the extent to which

the articles influenced the current research projects that the participants were undertaking.

Participant 7 claimed, “I think I would give something a ‘seven’ if it actually changed what I was doing in my study, influenced me in that way.” Below is an example of her use of the criterion.

I gave it a 7. A critical article ...first well-controlled study that shows that treatment I'm interested may be better than current treatment of choice. Lots of questions in discussion that my study can address. Lots of methods of assessment that I want to use now in my study.

Participant 2 also believed that “Influenced My Study” is an important criterion at Stage 2.

I had an idea for the design of the study before I went into this. And what this lit search did, what all these articles did was honing the idea and tighten it. Didn't significantly alter the design, but added many, many components to it.

Abstracts and their full-text counterparts. Participant 2 held strong opinions on how accurately the abstracts that she read represented the content of the full-text articles. For her selection, she indicated that there were quite a number of items whose abstracts were very deceptive.

Abstracts... are dry; they are not interesting. They don't tend to be that well organized. They are flip out after the paper was written. And they are just, several times they were really, really, really bad. The George Washington Crile one the abstract wasn't interesting at all, it was very flat. The paper was just amazing, very engaging, very well written and well sought out. So I think abstracts are very deceptive in that way.

...A lot of these were very deceptive. Look the abstracts look beautiful and the article sucked. Or like this one was like the abstract

was just long you couldn't see anything, and it was the best article of the bunch.

Participant 2 further pointed out that the abstracts she read could be grouped into three categories. The first type includes the ones that “overvalued” full-texts; the second group was the ones that “undervalue” the articles. There were also some items that she thought were “great,” and this group of abstracts was, in her own words, “right on.”

Participant 2's observations on the possible deceptiveness of abstracts echoes with the participants' comments in the experiment reported in Chapter 5 on the quality of the abstracts that they read. It is worth noting that the participants involved in the dissertation projects repeatedly indicated the discrepancy between an abstract and the full-text article the abstract is supposed to represent. This challenges the representational quality of abstracts, and suggests a need for current indexing and abstracting services to improve their sense of what people are looking for in such representations. It would be useful for future research to consider what makes an abstract a good representation of its full-text counterpart.

Consensus between the Use of Criteria and Perceptions of Criteria

It is interesting to compare the participants' actual use of criteria with their perceptions of the importance of the criteria used. Participants' actual use of criteria was reported in earlier sections. Table 6.6 and Figure 6.3 contain the micro level report; Table 6.10 and Figure 6.7 describe the use patterns on a macro level.

According to participants' own perceptions, “Topicality” was the single most important criterion for both Stage 1 and 2. As reflected in Figure 6.2, “Topicality

Relatedness” and “Topical Focus” were the two most frequently used criteria for both stages across all participants. From Figure 6.6, we can see that with the exception of *Research Structure* at Stage 2, *Topicality* has a much higher frequency rate at both stages than the rest of the criteria classes. It seems reasonable to conclude that the participants’ perceptions of the importance of topicality largely agree with what they actually did during the document evaluation processes.

Several participants believed that some secondary criteria, such as “Journal,” “Author,” and “Publication Date” all mattered more at Stage 1 but less so at Stage 2. Table 6.6 and Figure 6.3 show that in this respect participants’ perceptions also agree with their actual use patterns. The total relative frequencies of the three criteria all decreased at Stage 2, with “Journal” by 6%, “Author” by 5% and “Publication Date” by 3%. Although the values of the differences seem to be small, it is important to keep in mind that the maximum difference for all criteria commonly used by the nine participants across the two stages is 22%. On the macro level, these three criteria all belong to the class *Source Value*. As a group, *Source Value* decreased 8% at Stage 2, which, is also consistent with the participants’ reflections.

The participants appreciated the criterion “Newness” at Stage 2. On the micro level, this particular criterion increased by 5% at Stage 2. This matches the participants’ comments, although perhaps not as strongly as the participants had indicated. On the macro level, the criterion “Newness” belongs to the category

Cognitive State, which increased by 3% at Stage 2. This also is consistent with participants' reflections.

What participants' called "Writing Style and Presentation of Information," is similar to "Quality and Value" and "Readability." While the frequency rate for "Readability" remained the same across the two stages, "Quality and Value" increased by 8% at Stage 2. Participants believed that "Writing style" is more important for Stage 2, and this agrees with their actual use of the criterion. On the macro level, *Quality of Information* increased by 5%. Even though this is not a big increase, it to some extent confirms the participants' indication of what they actually did.

Crucial criteria for Stage 2 for most of the participants in this study were "Influenced My Study" and "Link to My Study." These criteria were unique to Stage 2. "Influenced My Study" had a relative frequency of 5% for Participant 7, but only 1% across all nine participants. "Link to My Study," on the other hand, increased by 8% at Stage 2. On the macro level, *My Study* had a great deal of change, increasing by 17% at Stage 2. On this level, participants' perceptions also agree with their actual use of criteria.

Overall, there was a good deal of agreement between participants' perceptions on the use of criteria and their actual use of criteria. The only minor discrepancy is with the criterion "Newness." The majority of participants discussed the important role of such a criterion at Stage 2, in the actual use, the increasing rate ($d = 5\%$) seems not as strong as how it was expressed by the participants. This

suggests that the importance of a criterion may not necessarily be reflected by the frequency of use. Some criteria may be employed occasionally, yet they matter very much in the few countable decisions. Another possible explanation is that the criterion “Newness” may be confounded with other criteria such as “Publication Date,” “Add My Knowledge,” and “Understandability,” causing the trend of change appeared less strongly than it would be.

Usefulness Rating and Definitions of Usefulness

On the document evaluation sheets, the participants were instructed to rate each of the articles read according to its usefulness. The evaluation sheets intentionally did not provide a definition of usefulness. During the post document evaluation interview, Participant 1 suggested that it might be helpful to clearly define the term “usefulness.” According to him, the notion “usefulness” here could carry two connotations --“utility in the sense of the particular project in question versus utility in the sense that, maybe the other projects you are already interested in, or perhaps even more general, their knowledge as scientists.”

Given Participant 1’s comments, I asked all the participants to define usefulness during the post document evaluation interviews. Five participants explicitly indicated that they considered usefulness in relation to the particular project they were undertaking at the time. Participant 8 said that she viewed usefulness as “primarily useful to her project,” though some articles got higher ratings because they were interesting to her, or they were helpful beyond the paper. Participant 1 also indicated that there was one article that “was extraordinary, and it was useful for this project, but particularly it was useful for honing my thinking

regarding my area of interest of self-regulation much more generally.” So he concluded “I am not sure whether I was using, is this generally useful to me as a scientist, or is this useful for this ego-depletion project I intended to do. I think I was sloppy about it.”

Participant 3 pointed out as she evaluated the articles that she placed about 60% of the value of each article on its relevance to the current project, and 40% on its relevance to future research. On the other hand, Participant 7 declared that her usefulness rating was based on “how useful for the actual execution of the study.”

Below is the exchange between the participant and the researcher on this matter:

Participant 7: Some of these helped my thinking as far as developing actual protocol even though I didn't include them in the proposal. Even though I didn't cite specific study in my proposal it still helped me think about what things I want to change in the way I designed my study.

Researcher: So they will be still high in the usefulness rating.

Participant 7: That's sort of how I defined usefulness, I think. How much it helped me in continue to design my study. So I didn't, in some cases, I did think it's useful to me as a researcher in general, I would say too bad, I would rate it low and say why it didn't help me with the study, and put in the parenthesis but it's good to know for me.

Overall, a majority of the participants rated the usefulness of their articles according to how useful the articles were to their projects. One thing that needs to be pointed out is that across participants usefulness had the greatest change from Stage 1 to 2. The frequency rate increased by 22% at Stage 2. This is very likely an influence of the document evaluation sheets for which participants were instructed to rate items according to usefulness. Participants could have been biased in their

approach to usefulness at Stage 2. Nevertheless, when explaining in detail how they defined usefulness, four participants specifically stated that usefulness incorporates topical relevance. Participant 5, in particular, claimed that usefulness means how well the articles help him understand one or more of his four topical themes.

Change in Criteria Use and Rankings of Criteria Importance

During post document evaluation interviews, participants were asked whether they believe that their use of criteria had changed as they moved from selecting bibliographic records to selecting actual articles. Following that, they were asked to rank the criteria they employed according to the importance of the criteria. Table 6.12 below provides an overview of participants' responses to these two issues. Notice that Participant 7 believed that "Topicality" contains two sub-elements, "About my topic" and "Population." For "Credibility," she suggested three sub-criteria, "Author," "Quality," and "Journal."

Table 6.12
Participants' Views on Changing and Ranking of Criteria

Participant No.	Change/ No Change	Criteria Ranking Stage 1	Criteria Ranking Stage 2
1	Yes	1. Topicality 1. Journal 1. Author 4. New Methods 5. Theoretical Newness	1. Topicality 2. New Methods 3. Theoretical Newness 4. Author 5. Journal
2	Yes	1. Types of Article 1. Topicality 3. Procedure 4. Population 5. Language	1. Newness 2. Design 3. Procedure 4. Writing Style and Presentation of Information 5. Recency 6. Support my ideas 7. Pursuit of Details
3	No	1. Topicality 2. Evidence of Effect 3. Specificity 4. Population (age group and ethnicity) 5. Method and Analysis	
4	Yes	1. Domain 2. Author 3. Specificity 4. Journal	1. Domain 2. Specificity 3. Have data 4. Author 5. Journal
5	Yes	1. Topical Focus 2. Population 3. Author 4. Journal 5. Recency 6. Cited in the preliminary paper	1. Topical Focus 2. Use of Statistics Method 3. Population 4. Construct 5. Writing Style and Definition of Terminology 6. Author 7. Journal 8. Mentioned by coauthor 9. Cited in the preliminary paper
6	No	1. Topicality 2. Design 3. Statistic Analysis	

Table 6.12
Participants' Views on Changing and Ranking of Criteria (Cont.)

Participant No.	Change/ No Change	Criteria Ranking Stage 1	Criteria Ranking Stage 2
7	No	1. Topical relevance About my topic Population 1. Methodology 1. Newness 4. Credibility Author Quality Journal	
8	Yes	1. Topical focus 2. Procedure and sample size 3. Types of article 4. Recency 5. Scope 6. Author	1. Topical focus 2. Types of article 3. Domain 4. Procedure and sample size 5. Recency 6. Author
9	Yes	1. Topicality 2. Journal 3. Author 4. Language	1. Topicality 2. Newness 3. Author

Six out of nine participants believed that there were changes in their use of criteria across the two stages. The other three indicated that they were employing the same criteria for selecting records and for selecting articles.

Despite the differing perceptions on the use of criteria, *Topicality* was viewed as the single most important criterion for document selection at both stages. All nine participants ranked topicality at the top; even people who claimed change in the use of criteria did not think that the rankings of "Topicality" changed over the stages. This group of people also seemed to highly value issues related to *Research Structure*. Criteria such as "Design," "Methods," "Procedure," "Population," "Analysis," "Have Data," "Sample Size" were ranked in most cases as a set of rather

important criteria, second only to “Topicality.” Participants placed more emphasis on this group of criteria at Stage 2.

What changed the most is the ranking of *Source Value* related criteria, such as “Author,” “Journal,” and “Recency.” Four out of six participants who suggested change in the use of criteria gave this group of criteria a reduced ranking at Stage 2.

In the meantime, four out of nine participants singled out the criterion “Newness,” and they gave this particular criterion a high ranking: first for two participants and second for the other two. However, as reflected in the actual use, the relative frequency for this individual criterion was 1% for Stage 1 and 6% for Stage 2. The increment in the use does not seem to be extraordinary large. From this, an interpretation may be made: Criteria differ by the nature of usage. Some criteria such as “Topicality” and “Research Structure” are essential for relevance evaluations; they tend to be frequently used at both stages. Criteria such as “Author” and “Journal” are more direct and easily employed. Other criteria such as “Newness” and “Inspirational” are more subtle, subjective and individually oriented, and therefore were not commonly used by participants at all times. However, this by no means suggests that participants considered these criteria to be unimportant.

Discussion

Many issues have already been touched upon during the earlier presentation of the results. In this section, I will first recapitulate the results based on the relative

frequency across the two stages and participants' recollections on their use of criteria. Next, I present two classification systems of criteria. In the end, I briefly examine the possibility of building a taxonomy of criteria that integrates the two systems. An attempt was made to classify some of the criteria used by the participants in this study according to the principles of the taxonomy.

The Most Frequently Used Criteria

Across the nine participants, "Topical Relatedness" was the most frequently used criterion at Stage 1, whereas "Topical Focus" was the most frequently used criterion at Stage 2. This transition seems to be quite reasonable because at Stage 1 people tend to have a broader standard for topicality and include items that appear to be generally related to their topics. When reading the actual articles at Stage 2, attention is on the central issues with which these articles have dealt. "Topical Focus" is a criterion that describes the specifics of the topic, and naturally it would be used more often at Stage 2 because full-text articles provide details about the topic. As noted by Khulthau (1993), people are generally more focused as they move from early exploration to the later stages of their information searching processes.

At Stage 1, other frequently used criteria include "Topical Focus," "Interestingness," and "Author." In addition to "Topical Focus," "Topical Relatedness," "Usefulness," "Design," "Interestingness," and "Results" were the most frequently used criteria at Stage 2. As participants read full-text articles, they not only were more discriminating in evaluating the topicality of the documents

and the interestingness of the documents, but also added other elements into consideration such as the utility of the document, the design and result of the study.

“Interestingness” was among the most frequently used criteria at both stages. This criterion has special properties, and it seems to hold an unpredictable relationship with topical relevance and usefulness. In evaluating a record, Participant 5 commented,

It's not very relevant, it's just using the circumplex as a way to measure or establish the construct validity when they are assessing an instrument questionnaire. Looks interesting, but not really relevant.

An item may be interesting but not relevant. In the same vein, a document can be viewed as interesting or intellectually exciting but not useful to the specific tasks that the participants had at hand. Participant 1 made such a comment regarding the articles he read: “in fact, even the ones that I rated low in usefulness, the majority of them were very interesting.” He further stated, “I got a lot from reading these articles.”

“Interestingness” was used frequently at both stages, and the use of the criterion was also very stable. Every participant used the criterion in each of the two stages except Participant 4, who did not use the criterion for his Stage 2 evaluation. Across the participants, the frequency rate for “Interestingness” only decreased slightly ($d = -2\%$) at Stage 2. This suggests that “Interestingness” is a criterion that may be employed by all the people at all times.

On the other hand, “Interestingness” may be used to describe a variety of aspects of the documents. In other words, “Interestingness” carries a rather flexible

set of connotations depending on the document reviewed. An item could be viewed as “interesting” because it contains interesting concepts, or theories, or propositions, or methods, or results, or ways of expression, or even some unusual facts.

To summarize, “Interestingness” holds its own properties, and its use seems to be independent from other major criteria. It is used regardless of the stage in the evaluation process. It may be used to describe various characteristics of a document. One of the potential research questions may be to investigate the association between interestingness and topicality on the one hand and the relationship between interestingness and utility (usefulness) on the other hand.

The Most Changed Criteria

The criteria that shifted the most in frequency of use from Stage 1 to 2 include “Usefulness,” “Design” and “Topical Relatedness.” The first two criteria increased in frequency rates at Stage 2, whereas at the same stage the use of the third criteria was decreased.

As discussed earlier, the high mentioning rate of “Usefulness” at Stage 2 may have been encouraged by the fact that participants were asked to rate the articles based on usefulness. If we suppose that without this bias, “Usefulness” was still employed more frequently at Stage 2, it then suggests that participants’ attention was more on the utility aspect of the documents at Stage 2 than at Stage 1. It may also suggest that the primary selection criteria switched from topically centered heuristics for record evaluation to more utility oriented principles for full-text evaluation.

While “Design” was mentioned much more frequently at Stage 2, “Topical Relatedness” was used much less frequently. These changes also seem reasonable. When reading actual articles, participants could have started to look for things that are directly related to the projects at hand. One of these things is the research design, i.e., whether the design employed in the study could help them to structure or restructure their ongoing research projects. At this point, participants had already used “Topical Relatedness” to filter out unwanted records, and hence most of the articles they read already would be related to their topics. As a result, participants paid much less attention to the topical relatedness of the articles, and were concentrated more on the details of the design of the study.

Other criteria that were used more frequently at Stage 2 include “Results,” “Quality and Value,” and “Link to My Study.” “Journal,” “Is About” and “Author” were used less frequently at Stage 2. As discussed earlier, participants used “Journal” and “Author” mostly at Stage 1 to eliminate unwanted items, at Stage 2 all the articles that participants had selected had already fulfilled their requirements, so these criteria became less of a concern at that time and were used less frequently.

Most Frequently Used Criteria Classes and Most Changed Classes

On the macro level, *Topicality* was the most frequently used class at Stage 1. At Stage 2, the top category became *Research Structure*. Participants applied *Topicality* more often to select bibliographic records, however, after they read full-text articles, they based their selection decisions more on research related aspects

such as design, procedure, population, variables, and results. One of the reasons that this group of participants was strongly interested in research mechanisms is their academic status and research backgrounds. The field of psychology is a discipline with a long and well-established research tradition. Psychology also has many well-developed theories and a highly mature structure for empirical research. Most of the participants had in-depth experiences in conducting experimental studies, and they were involved as the participants in this research because they really needed some literature to help them to develop their own research. While at the first stage, participants selected records mostly based on topicality, at Stage 2, they were really looking at specifics in research structure and making connections with regard to how useful the articles were to their interests.

Other evidence of this is in the use of the class *My Study*, which increased by 17% at Stage 2. This increase suggests that participants focused more on whether the document read would enhance their research projects. If measured by the change in ranked positions, *My Study* was the class that had the greatest positive change ($d_r = 2$). On the other hand, the most negatively changed class in the rankings is *Source Value*. Participants employed *Source Value* as the third most frequently used class at Stage 1, but at Stage 2 it became the seventh, dropping down from a set of eight by four units. Earlier analysis provided rationale for why *Source Value* as a criteria class would be used less at Stage 2.

Process Model Tested

The dissertation study started with a proposition based on a Process Model claiming that the criteria used for selecting bibliographic records would be different from the criteria used for the full-text evaluation. The Process Model further states that at Stage 1 users center their attention on *Topicality*, but as they move to Stage 2, they place emphasis on *Cognitive* aspects and/or situational elements.

The results from this study showed that the use of *Topicality* did decrease at Stage 2 for this group of participants and the character of topicality related criteria changed from Stage 1 to 2. The results also show that the use of *Cognitive State* increased slightly at Stage 2, although not as much as other classes such as *Research Structure*, *Utility*, *My Study*, or *Quality of Information*.

Taxonomies of Criteria

The report of the laboratory results in the previous chapter led to the observation that classifying criteria semantically may not achieve an accurate and consistent grouping. The four factor solutions generated by Factor Analysis showed a trend in criteria to cluster by their general nature of being either objective or subjective. The idea of classifying criteria according to their nature was reinforced in this study during the discussions with the nine participants. Furthermore, from participants' verbal and written comments, it appeared that in addition to the distinction between subjective and objective, the criteria may also be grouped by their functionality. The functionality of a given criterion is determined by whether the criterion is used as an essential/primary element or a supportive/secondary

factor for the decision making. I therefore propose two classification systems of criteria: a) primary criteria versus secondary criteria, and b) objective criteria versus subjective criteria.

Primary versus Secondary. Criteria can be grouped into different types based on what role they play or what they contribute to people's relevance judgments and document selection decisions. Some criteria were used by participants as the primary reasons for their decisions; others were used as a qualifier to support an otherwise undecided position. Depending on people's information needs, their search tasks, their knowledge backgrounds, and their personal characteristics, a criterion may be treated as the primary criterion for a particular person with a specific task at a particular point in time; it may be considered as a secondary criterion at another time for the same person who has a different task or who is at a different stage of a task. It might be helpful for an IR system to help its users to analyze the nature of a given task and select and prioritize their criteria, to retrieve documents that better fit their situational needs. Upon differentiating the Primary Criteria from the Secondary Criteria, the IR systems equipped with filtering capabilities may use Secondary Criteria to reduce the retrieval set and help people focus their cognitive resources on evaluation using Primary Criteria. There is more detailed discussion about the implication for system design in the next chapter.

“Topicality” is typically a *Primary Criterion*, since it is the basis for most of the document evaluations and selections. In this study, “Topicality” (“Topical

Relatedness,” “Topical Focus”) was used most frequently at both stages. And with little exception, the participants believed that “Topicality” was the most important among all other criteria used.

Secondary Criteria normally do not constitute central decision-making factors. Typically, “Author,” “Journal,” as well as “Recency,” would not be the stand-alone factors for relevance decisions, unless it is for a known-item search. In this study, the participants used these three criteria as secondary criteria to aid their choices when the topical information of the item was ambiguous or not quite complete. These three criteria were applied much more frequently when records did not include abstracts.

As participants commented on their evaluation processes either during or after the actual evaluation, they believed that they typically used “Author,” “Journal,” and “Recency (Publication Date)” as “watching” criteria or “rules of thumb.” These criteria, especially the first two, can help them to make an “educated guess” about the relevance of the documents. All participants recognized the value of these three criteria. Nonetheless, except for Participant 1, other participants did not rank the criteria as the highest rated criteria either for their record or full-text selection processes. For people who believed that there were changes in the importance of criteria from Stage 1 to 2, they all indicated that the values of “Author,” “Journal,” or “Publication Date” were decreased to some extent at Stage 2. In other words, these secondary criteria became much less important in the later stage of document selection.

Objective versus Subjective. A second classification system divides criteria into two groups: Objective Criteria versus Subjective Criteria. Objective Criteria are those that describe the characteristics of a given document; Subjective Criteria are the ones that relate to the users' situation. Objective Criteria are more likely to be employed when reasoning is based on the properties of the documents. Subjective Criteria are rationales formed from the perspective of the user, as in whether the user "likes" the document personally, whether the document influence a user's knowledge structure, or whether the documents have utility and satisfy needs.

Of the criteria listed in Table 6.6, most of the criteria under *Source Value* are Objective Criteria, as are most of the criteria under *Topicality*, *Research Structure* and *Quality of Information*. Most of the criteria under *Cognitive State*, *Affective Aspect*, *Utility*, and *My Study* are Subjective Criteria.

In this study, Objective Criteria (average freq. Stage 1 = 36%; average freq. Stage 2 = 40%) were used more frequently at both stages than were Subjective Criteria (average freq. Stage 1 = 9.3%; average freq. Stage 2 = 19%). However, Subjective Criteria ($d_r = 9.5\%$) had a greater rate of increase than Objective Criteria ($d_r = 4.5\%$). This suggests that although Objective Criteria were important and frequently employed throughout the stages of the document evaluation processes, Subjective Criteria increased greatly from Stage 1 to Stage 2.

An integrated taxonomy. A next step with these two systems of criteria is to combine the two into an integrated taxonomy of criteria. Four categories of criteria constitute this taxonomy: Primary Objective Criteria, Primary Subjective Criteria,

Secondary Objective Criteria, and Secondary Subjective Criteria. Primary Objective Criteria are the criteria that are essential to users' evaluation decisions and are related to the documents being reviewed. Secondary Objective Criteria are the criteria that support users' evaluation decisions and are related to the documents being reviewed. Primary Subjective Criteria are the criteria that are essential to users' evaluation decisions and are related to the users' situations. Secondary Subjective Criteria are the criteria that support the users' evaluation decisions and are related to the users' situations.

As discussed earlier, what makes a criterion primary or secondary is dependent on a given task, the evaluation stage, and the individual. Thus, it is in some ways arbitrary to group all the criteria used by the participants in this study to the four categories of the taxonomy. Nonetheless, Table 6.13 below illustrates an attempt to categorize criteria and establish the possible instances for the taxonomy, based on the results of the frequency counts and participants' own perceptions of the use of criteria. Readers are warned that the following classification only serves as a possible grouping for about one quarter of the criteria used by the participants involved in this study. It may vary for a specific participant, and it is not generalizable to other research populations. The purpose of this table is to give an idea of what each category in the taxonomy might mean in the case of this study.

Table 6.13
A New Taxonomy of Criteria

Primary Objective	Primary Subjective	Secondary Objective	Secondary Subjective
Domain	Add My Knowledge	Author	Interestingness
Is About	Informativeness	Cited Frequently	Affective
Topical Focus	Inspirational	Journal	
Topical Relatedness	Newness	Length of Article	
Concepts	Support My View	Publication Date	
Conclusion	Helpfulness	Reference	
Data	Usefulness		
Design	Influenced My Study		
Evidence of Effect	Is What I Want		
Method	Justification of My		
Nature of the Study	Study		
Population	Link to My Study		
Procedure	Similar to What I Do		
Results			
Statistic Analysis			
Techniques			
Theoretical Model			
Variable and			
Constructs			
Completeness			
Scope			
Quality and Value			
Article Type			

This new taxonomy of criteria not only has impacts to the conceptual development of relevance study, it also provides useful ideas to the design of bibliographic retrieval systems. These ideas are discussed in full detail in the next chapter.

Summary

In this chapter, major results of the naturalistic study are presented. On the micro level, it was found that “Topical Relatedness” was used most frequently at Stage 1, while at Stage 2 “Topical Focus” was used most frequently. The use of the criterion “Usefulness” changed the most moving from Stage 1 to Stage 2, although it was pointed out that the change might be biased because of the evaluation instrument. The second greatest change included the criterion “Design,” which was used more frequently at Stage 2, and the criterion “Topical Relatedness,” which was used less frequently at Stage 2. On the macro level, it was found that the most frequently used categories were *Topicality* at Stage 1, and *Research Structure* at Stage 2. The highest change was with *Research Structure*, increased greatly at Stage 2. *My Study* had the second highest change, and was used more often at Stage 2. Both *Topicality* and *Source Value* had higher frequency rates at Stage 1 than at Stage 2.

Participants’ reflections on the use of criteria served as one important component for data analysis. Participants believed that “Topicality” was the single most important criterion for both Stage 1 and Stage 2. They also promoted the value of the criterion “Newness” especially for Stage 2. They claimed that “Newness” should be an essential and overriding criterion for people when they evaluate full-text documents. Participants’ perceptions were mostly consistent with their actual use patterns.

The findings of the study seemed to confirm the hypothesis projected by the Process Model of Relevance Judgments, as explicated in Chapter 3. However, it is

argued that the model may be further extended to include more categories, and it may be reconstructed according to a taxonomy that contrasts primary/secondary and objective/subjective criteria. More discussion of the taxonomy and modification of the Process Model will be presented in the next chapter.

Recall that this study focuses on nine participants with the common background of advanced study in Psychology and enrollment in a meta-analysis course. The intent of the study is, accordingly, not to generalize to a broader population, but to map the patterns of change in criteria use of the participants as they evaluate documents for their course products. Underlying this intensive mapping is an interest in developing and refining our understanding of criteria that are employed in document evaluation and elaborating a Process Model of document evaluation criteria and their change. This study, thus, provides grounding for future theorizing. This theorizing will need to be shaped by other mappings from other situations and subject domains such as is given in Chapter 5.

Chapter 7

CONCLUSIONS AND IMPLICATIONS

Introduction

The purpose of this study was to empirically investigate the criteria that people employ when they evaluate bibliographic records (Stage 1) and then full-text articles (Stage 2). Two studies with different research designs were conducted to support the investigation. The two level, micro- and macro- level, analysis of the data enabled not only an intensive examination of the use of individual criteria but also a more general understanding of the dimensional movement of such evaluation criteria in the form of classes of criteria.

The research reported here contributes to the knowledge base supporting the development of information retrieval systems and an associated theory of relevance in three ways.

1. The findings helped to reach a multi-level understanding of the dynamics of criteria usage during the process of document evaluation. The results also helped to augment the theory of relevance by offering a research based taxonomy of criteria.

2. The study established research protocols that applied both naturalistic and laboratory approaches to the investigation of users' criteria. These protocols challenge the domination of a single approach in the empirical design of studies of relevance criteria. The experience of their applications suggests that laboratory experiments, if designed properly, may serve as an alternative approach for the investigation of relevance criteria. That is, a controlled, laboratory design could be used to test research hypotheses that predict the specific use patterns of relevance criteria.
3. As exploratory research, this study has identified a set of research issues that provide a basis for future research and theory development.

This chapter starts with a statement of conclusions drawn from the laboratory and naturalistic studies, which were summarized in Chapters 5 and 6 respectively. The next sections focus on the implications of the results from several points of view: understanding of the document evaluation process, theoretical development in the study of users' criteria, the design of bibliographic retrieval systems, research methodology, and future research.

Conclusions Based on the Results

Research Question 1: Use of Individual Criteria across Stages of Document

Evaluation

In response to Research Question 1, the study identified the criteria that were rated as the most important by the laboratory experiment participants or used most frequently by the naturalistic study participants at each of the two stages. The study also described change patterns in the usage of individual criteria across the evaluation of bibliographic records (Stage 1) and the evaluation of full-text articles (Stage 2).

In terms of the ratings of importance, the laboratory experiment acquired participants' importance ratings of the 15 criteria. The naturalistic study also collected participants' perceptions on the importance of the criteria that they applied during their natural processes of document evaluation. The participants in the laboratory experiment rated "Understandable, not too technically complex", i.e. the criterion "Understandability" as the most important criterion for Stage 1. "Discuss Y2K and its social effect" was rated as the most important criterion for Stage 2. "Discuss Y2K and its social effect" is semantically equivalent to the criterion "Topical Focus." In the naturalistic study, on the other hand, most of the participants believed that "Topical Focus" was the single most important factor for both stages, with the exception of one participant (Participant 2), who ranked "Newness" as the number one criterion for full-text evaluation. In summary, across the two studies, "Understandability," "Topical Focus," and "Newness" were perceived to be important for the two stages of document evaluation.

In terms of frequency of use, the naturalistic study found that the most frequently used criterion across the nine participants was “Topical Relatedness” for Stage 1 and “Topical Focus” for Stage 2. Evidently, topicality was used most frequently across the stages, with the shift from a broad, general standard of “Topical Relatedness” at Stage 1 to a focused, specified standard of “Topical Focus” at Stage 2.

In terms of change in importance ratings, the laboratory experiment found that the most changed criterion was “Issues are real and important” and that the least changed criterion was “Fresh and unique approach.” Several participants in the naturalistic study claimed that the criterion “Newness” should have increased importance at Stage 2, whereas the importance for criteria such as “Author,” “Journal,” and “Recency” should have decreased for the full-text evaluation.

An attempt to cross tabulate the results of the laboratory experiment and the naturalistic study on a micro individual criteria level is presented in Table 7.1. The table lists the change pattern ($d = \text{measure of Stage 2} - \text{measure of Stage 1}$) for some of the criteria that are common in the two studies. Note that the two studies used different measurements. The experimental study measured the participants’ ratings of importance of criteria while the naturalistic study measured the frequencies in the participants’ actual use of criteria. The measuring units are thus not directly comparable.

Table 7.1
Patterns of Change in the Use of Criteria, Naturalistic versus Laboratory

Criteria	Naturalistic	Laboratory
-----------------	---------------------	-------------------

Background Information	No change ($d = 0\%$)	Increase ⁴ ($d = 0.33$)
Clarity	No change ($d = 0\%$)	Increase ($d = 0.05$)
Data	Increase ($d = 5\%$)	Increase ⁵ ($d = 0.24$)
Importance	Increase ($d = 2\%$)	Increase ($d = 0.39$)
Interestingness	Decrease ($d = -2\%$)	Increase ($d = 0.13$)
Newness	Increase ($d = 5\%$)	Increase ($d = 0.13$)
Recency	Decrease ($d = -3\%$)	Decrease ($d = -0.14$)
Topical Focus	Increase ($d = 6\%$)	Increase ⁶ ($d = 0.33$)
Topical Relatedness	Decrease ($d = -11\%$)	Increase ⁷ ($d = 0.06$)
Understandability	No change ($d = 0\%$)	Decrease ($d = -0.13$)
Usefulness	Increase ($d = 22\%$)	N/A

Note: d = rate of Stage 2 – rate of Stage 1

The Process Model predicts that the criteria related to *Topicality* demonstrate a decreasing trend from Stage 1 to Stage 2. In this study, “Topical Focus” was rated more important at Stage 2 by the laboratory participants and used more often by the naturalistic participants at Stage 2. “Topical Relatedness” was used less often at Stage 2 by the naturalistic participants but was rated a little more important at Stage 2 in the laboratory experiment.

The Process Model predicts an increasing pattern for the criteria that are connected with users’ *Cognitive State*. The study found that the criterion “Newness” was both used more frequently at Stage 2 by the naturalistic participants and rated more important at Stage 2 by the laboratory participants and some of the naturalistic

⁴ Criterion “Cover Y2K origin and causes” is considered as “Background Information.”

⁵ Criterion “Factual information and actual data” is considered as “Data.”

⁶ Criterion “Discuss Y2K and its social effect” is considered as “Topical Focus.”

⁷ Criterion “Provide definition of Y2K” is considered as “Topical Relatedness.”

participants. This supports Boyce's proposition that "novelty," as an element of "informativeness," receives more attention at the later stage of judgment. On the other hand, "Understandability" did not change for the naturalistic study and it was rated as less important by the laboratory participants. This is inconsistent with Boyce's theory that people focus more on "Understandability," another element of "informativeness" at the later stage of evaluation.

"Importance" increased at Stage 2 both from the laboratory perspective of the importance rating and from the naturalistic perspective of use frequency.

"Interestingness" was used a little less often at Stage 2 in the naturalistic setting but was rated higher in importance by the laboratory participants at Stage 2. In summary, while several criteria related to users' cognition presented an increasing trend, a few others did not.

The Process Model also predicts an increasing pattern for the criteria related to *Utility*. While "Usefulness" was not measured in the laboratory experiment, the naturalistic study confirmed that the participants applied this criterion more frequently for full-text evaluation.

"Recency" demonstrated a decreasing trend both in the laboratory rating and the naturalistic use. This criterion was grouped under *Quality of Information* with the laboratory structure, but was grouped as *Source Value* under a finer naturalistic model. "Data" was treated as increasingly important and used more often at Stage 2. This criterion was under *Topicality* in the laboratory study but belonged to *Research Structure* in the naturalistic grouping. "Background Information" was rated

as more important at Stage 2 but presented no change when measured by the frequency of use. “Background Information” was considered as a property of *Topicality* in the laboratory study but was grouped within the *Quality of Information* class in the naturalistic context. The criterion “Clarity” showed no change as measured by the frequency rates in the naturalistic study, but was rated as more important by the participants in the experiment. Two participants from the naturalistic study also articulated a strong concern about writing style and clarity of the articles in their Stage 2 evaluation.

Overall, in the micro level examination of individual criteria, the Process Model was partially confirmed by the patterns of change for some of the criteria. Yet the model demonstrated certain limitations in describing the nature of change for several other criteria.

Research Question 2: Use of Classes of Criteria across Stages of Document Evaluation

In response to Research Question 2, the researcher applied two series of category structures for the analysis of data. For the laboratory data with the 15 predefined criteria, an a priori model that featured three classes of criteria was used. This a priori model was tested and challenged by the solutions generated through Factor Analysis. On the other hand, the naturalistic data provided a rich set of criteria, which also resulted in the expansion and modification of the a priori category structure. The classification resulting from the analysis of the naturalistic study has eight classes in total, including the three original classes.

In restructuring the categories, some criteria were grouped under different classes. For instance, “Cover Y2K origin and causes,” which is a criterion related to “Provides background information,” was thought to be appropriate as an element of *Topicality* for the laboratory experiment. However, in the naturalistic context, participants considered whether or not an item provided background information more as a feature of the quality of the paper. If a document contained background information, it was viewed as of better quality. Therefore, this criterion was ultimately categorized under *Quality of Information*. “Provides factual information and data” was considered as an attribute of *Topicality* in the a priori model for the laboratory experiment. In the naturalistic study, this particular criterion was used in the context when participants were looking for whether the document contained experimental data. It is therefore seen as addressing the research mechanism of the document and “Data” was, therefore, categorized into the class *Research Structure*. “Information up to date” was considered as part of the *Quality of Information* for the laboratory study, whereas “Publication Date” or “Recency” was grouped into the *Source Value* category. *Source Value* is an added category that was not present in the a priori structure.

In terms of the ratings of importance for the three classes of criteria, the laboratory participants rated *Quality of Information* as most important for Stage 1 and *Topicality* as most important for Stage 2. *Cognitive State* obtained the lowest ratings at both stages. In terms of the frequency of use, among the eight classes of criteria, the naturalistic participants used *Topicality* most frequently for Stage 1 and *Research*

Structure most frequently for Stage 2. The least frequently used criteria class was *Affective Aspect* for both stages.

In terms of change from Stage 1 to Stage 2, the importance ratings for *Topicality* changed the most, while *Quality of Information* changed the least. The most changed criteria class, as measured by the frequency of use, was *Research Structure*, which was used most frequently at Stage 2. The least changed criteria class was *Affective Aspect*, which was used evenly infrequently at both stages.

Table 7.2 contrasts the findings from both studies. Again, the measurement for naturalistic study is the frequency of use whereas laboratory experiment measured the ratings of importance. The two studies share three common classes, and the naturalistic study has an extended structure with five more classes of criteria.

Table 7.2
Patterns of Change in the Use of Criteria Classes, Naturalistic versus Laboratory

Criteria Classes	Naturalistic	Laboratory
Topicality	Decrease ($d = -14\%$)	Increase ($d = 0.22$)
Cognitive State	Increase ($d = 5\%$)	Increase ($d = 0.11$)
Quality of Information	Increase ($d = 7\%$)	Increase ($d = 0.08$)
Research Structure	Increase ($d = 35\%$)	N/A
Source Value	Decrease ($d = -8\%$)	N/A
Utility	Increase ($d = 18\%$)	N/A
My Study	Increase ($d = 17\%$)	N/A
Affective Aspect	No change ($d = 0\%$)	N/A

The Process Model predicts a decreasing pattern for *Topicality*. This was confirmed in the naturalistic study: the participants used *Topicality* less frequently for the full-text evaluation. However, the laboratory participants rated *Topicality* as

more important at Stage 2. Most of the participants in the naturalistic study believed that *Topicality* would be as important for Stage 2, and one participant believed that it was more important for Stage 2 evaluation. Overall, *Topicality* was perceived as an important dimension for both record evaluation and full-text evaluation, however, it was used less frequently for full-text evaluation in the naturalistic study.

The Process Model predicts an increasing trend for the dimension of *Cognitive State*. It was confirmed by the results of the study both from the perspective of the importance rating and from the perspective of frequency. The naturalistic participants employed *Cognitive State* more frequently for their full-text evaluation, while the laboratory participants rated *Cognitive State* as increasingly important for Stage 2.

The Process Model also predicts an increasing pattern for *Utility*. This class, although not included in the laboratory study, was indeed applied more frequently by the naturalistic participants. A consistently increasing pattern for the dimension of *Quality of Information* was also found. Participants used this class more often at Stage 2 and rated this class as more important at Stage 2.

The classes of criteria that were not included in the laboratory experiment design also present interesting patterns of change in the naturalistic context. *Research Structure* and *My Study* were used more frequently by the participants for their full-text evaluation, whereas *Source Value* was used more often for record

evaluation. *Affective Aspect* was found used infrequently and presented no change across the two stages.

Overall, on a macro level analysis of criteria classes, the Process Model was partially confirmed. While the naturalistic data roughly supports the propositions of the Process Model, the results of the laboratory experiment presented a different pattern of change for *Topicality*.

Implications for Understanding Relevance in Document Evaluation Process

Previous research has established that relevance is a multidimensional construct and users employ multiple criteria for their relevance judgments. It had been found that users applied criteria that are not only related to the topical aspects of the documents but also associated with their own situational and cognitive needs. Based on the evidence from the two empirical projects, this research confirmed that users employ multidimensional criteria for their document evaluations and that the use of criteria is dynamic and evolving over the stages of evaluation. Specifically the research found that as participants moved from Stage 1 (evaluation of bibliographic records) to Stage 2 (evaluation of full-text documents), their reasoning structures shifted as manifested through their reprioritization of the criteria and adjustment of focus. The naturalistic study revealed a specific pattern of change with participants relying mainly on *Topicality*, *Research Structure*, and *Source Value* for the record evaluation and *Research Structure*, *Topicality*, *Cognitive State*, *My Study*, and *Quality of Information* for the full-text evaluation. On the other hand, the

laboratory experiment showed a shift from use of *Quality of Information* and *Topicality* for Stage 1 to *Topicality* and *Cognitive State* for Stage 2. Both studies suggested that people do reprioritize and adjust their reasoning factors as they advance from one stage of the document evaluation to another.

This study opened a new area of research in the study of relevance by operationalizing change in the use of criteria at two concrete stages of the document evaluation process. Based on the ratings of importance and frequency of use, the research examined the direction of actual movements in people's reasoning structure from both a micro individual criteria perspective and a macro dimensions of criteria perspective. Previous research has not focused on the exact change in criteria usage across the two stages and on the two levels as specified in this study. This line of research will provide grounded data to enhance our understanding of dynamic nature of relevance judgments in relation to the processes of document evaluation and document selection.

The role of *Topicality* in the document evaluation process is worth further discussion. For a long time *Topicality* has been viewed as the basis of the system-oriented relevance (Schamber, et al., 1990). Various writers have argued that centering on topicality limits the judgment of relevance, and ignores the role of the searcher. For instance, Boyce (1982) asserted that "Topicality is operationally necessary but insufficient condition" for relevance judgments. He further explains,

If one is to have relevance, that is, if one is to have satisfaction of user's need, something else is required. Topicality may or may not be necessary for relevance. It is surely insufficient ... It is certainly clear

that the requestor is his final judgment makes use of criteria other than topicality. (p. 105-106)

With the realization that topicality is not the sole factor in document evaluation, much research has been conducted to elicit a comprehensive list of all the criteria that people use to make their relevance decisions. Empirical evidence suggests that people apply criteria other than *Topicality*. On the other hand, it has been repeatedly found that *Topicality* is one of the major criteria that users employed. The important role of *Topicality* should be fully recognized and not be downplayed. To use Boyce's terms, I think that even though *Topicality* as a factor by itself is insufficient, it is still a "necessary" condition for relevance judgment.

In fact, many studies have shown that *Topicality* was the most frequently used criterion for document evaluation. For example, Barry (1994) found that "Criteria pertaining to information content of the document" was mentioned most frequently by all respondents. Wang (1994) also concluded that "topicality was the most frequently mentioned criterion for all participants" (p. 179) during their evaluation of document surrogates.

Considering the nature and use of *Topicality* as a class of criteria, Bateman (1998b) made the following observations in her research:

- Topicality (aboutness) is usually highly situational and dependent on the user, his or her situation and the information problem (p. 147).
- It is difficult to identify criteria to measure this construct that are valid and reliable across users and information situations and problems (p. 147).

- It is likely that topicality interacts with many of the other criteria (p. 147).

The study here also examined the role of *Topicality* across the two stages of document evaluation, and found that *Topicality* functioned as a central factor throughout the document evaluation process, though the nature of topicality shifted. *Topicality* was rated as the most important criteria class for Stage 2 by laboratory participants. It was the single significantly changed factor according to the Hotelling T² analysis. It was also rated by the naturalistic participants as the most important element for both stages. *Topicality* was used most frequently at Stage 1 and second most frequently at Stage 2 by the naturalistic participants. The participants shifted their focus slightly from *Topicality* to *Research Structure* as they moved to full-text evaluation. Nonetheless, the results of this study confirmed that *Topicality* is not only a “necessary” condition but also an essential factor for document evaluation.

Having established the important role of *Topicality*, it appears that there is a need for more studies to investigate the properties and characteristics of this particular dimension of criteria. This researcher agrees with Bateman’s observation that contrary to the normal understanding, *Topicality* is a complex and subtle factor. It is highly situational dependent. As a class of criteria its criteria elements or attributes are difficult to identify. In addition, *Topicality* often interacts with many other criteria or criteria classes. Using the naturalistic study as an example, *Topicality* is closely associated with the elements under *Research Structure*, and it is supported by criteria such as “Author” or “Journal.” Nevertheless, it is important

both to the theoretical development of relevance and to the design of retrieval systems to investigate the specifics and the functionality of use of the *Topicality* dimension in users' document evaluation processes.

Implications for Theoretical Development

Previous research on relevance criteria has established the convention of classifying relevance criteria on the basis of their meanings. This meaning-oriented approach has caused inconsistent structuralizations of criteria because many criteria have multiple meanings and may be grouped into different categories depending on the context of the research, the nature of the tasks, and the participants' specific needs for information. This dissertation research also began with this thematic approach and established a three-class structure for the 15 criteria used in the laboratory study. During the data analysis of the naturalistic data, it was found that the researcher's original classification structure needed to be modified and expanded to reflect actual behavior.

Towards the end of the data analysis, the researcher realized that it would be a challenge to any research effort to attempt to group users' criteria under finer conceptual dimensions and still justify its validity. The diversified structuralizations of criteria resulted from the Factor Analysis of laboratory data and recategorization of criteria based on the naturalistic data indicate that the meaning of the individual criteria is highly dependent on topic, task and user. Consequently, consensus may be difficult to reach if the category system takes a thematically oriented or meaning-based approach. There would be debates

concerning the specific group membership of some individual criteria and the contrasts in the laboratory and naturalistic studies showed that some criteria were employed in different ways by undergraduates taking a psychology class and graduate students in psychology. The researcher hence saw a need to categorize the criteria not through content connotations or semantic identities of criteria but through their more general nature and functionality.

The researcher developed two systems of classification for criteria. The first system classifies criteria by their nature: Objective and Subjective. The criteria that are closely related to the characteristics of a *document* as an entity are deemed to be Objective, whereas the criteria that are closely associated with a *person's* interpretations are considered to be Subjective. The second classification system divides criteria according to their functionality: Primary and Secondary. Primary criteria are *essential* for relevance decision making, whereas Secondary criteria are used to *assist* the decision making. It is also noted that the functionality of a criterion is flexible. Whether a criterion is primary or secondary depends not only on the topic, the task, the individual, but also on the stage in the process. A criterion may work as a primary criterion at one stage but become secondary at another stage.

Taxonomy of criteria was built by integrating the two systems. It is thus proposed that relevance criteria may be categorized into the four classes as follows: Primary Objective, Primary Subjective, Secondary Objective and Secondary Subjective Criteria. Primary Objective Criteria are related to the *document* and are

essential to the judgments; Primary Subjective Criteria are related to the *user* and are *essential* to the judgments. Secondary Objective Criteria are related to the *document* and are *supportive* factors for decision making. Secondary Subjective Criteria are related to the *user* and are *supportive* factors for decision making.

One of the advantages of this taxonomy of criteria over conventional content-based approach is that it is less attached to specific situations of criteria usage and hence may be better at generalizing trends of movements or patterns of change across contexts. Too, by classifying criteria by their nature and functionality, this taxonomy is easily operable to describe specific situations of criteria usage. It avoids the fuzziness in grouping by meaning and is easily comprehensible to searchers and designers as well as researchers.

As a result, a revised Process Model was developed based on this taxonomy of criteria. The model is now not restricted to describe movements of criteria measured by individual thematic dimensions. This Process Model now predicts that as users progress from Stage 1 (record evaluation) to Stage 2 (full-text evaluation), the criteria that they employ shift from a relatively strong objective orientation to an added subjective emphasis.

Implication for System Design

Many of the findings of this research have implications for building a bibliographic retrieval system that better satisfies users' needs. In the following paragraphs, I will concentrate on two findings in particular.

This study verified that relevance is dynamic in nature. More importantly, the study found that the participants reprioritized their reasoning factors as they move from Stage 1 to Stage 2. This finding suggests requirements in the design of bibliographic retrieval systems. Such a system should be flexible enough to accommodate the complexity of users' needs and adjust to the evolving nature of users' reasoning. Such a system would incorporate users' instant feedback on the impact of various criteria to their choices. I further envision that the system include the following three basic features:

- It allows multiple field search. Similar to the search on Dialog, including search capability on both basic indexed fields and additional indexed fields.
- It provides a relevance feedback mechanism.
- It is interactive in nature, equipped with machine learning capacity.

As an added component to relevance feedback, the system would provide an interface that allows people to specify the criteria that they view as important for their selection purposes. One possibility is to incorporate this component as a part of the whole search interface. This holistic interface would initially provide users with a comprehensive list of criteria and prompt the user to select a set of criteria that they wish to use in evaluating documents. The system then would ask the user

to rank criteria by their importance or assign weights to the criteria specified. Users' input on the weights of the criteria would be integrated with their query statements and be used to support a field-based search. To use Participant 8's topic as an example, suppose that she is searching the system for articles on intervention of women with coronary heart disease (CHD) and that she has a very specific search need. On the search interface, she specifies *query* term as intervention and heart disease, and she also specifies *population* to be women, *age* to be above fifty, and *document type* as clinical trial (assume that all the italics are searchable fields in the system). She then specifies in her criteria list that *document type* is the most important criterion. Then the system would arrange the retrieved documents in order by *document type* with clinical trial displayed first. Since the system is interactive, the user would be able to change her selection of criteria or adjust her ranking of criteria at any time during the search.

With Objective Criteria, the retrieval algorithm may be relatively easier to build. It is more difficult to construct an algorithm for Subjective Criteria. Since Subjective Criteria are individual specific, the system needs to learn from a user's evaluation of documents in order to profile the criteria that are uniquely oriented to that user. One possibility for providing this option is that for each item retrieved, the system prompts the user with a set of questions and records and learns from the responses of the user. Using Participant 1 as an example, suppose that he is searching on the system, and he indicates that "Theoretical Newness" and "Inspirational" are the two most important criteria for his selection. For the first

few iterations, along with the presentation of each item, the system would inquire the user to indicate how satisfied he is with the “Theoretical Newness” and “Inspirational” of a given document. Using the user’s feedback in combination with the specifics of documents content, the system would retrieve documents that are similar to the documents for which the user has expressed high satisfaction.

The second important implication for the system design is based on observing participants’ reaction to the representational quality of abstracts. Participants from both laboratory and naturalistic studies indicated that some abstracts were very deceptive and did not accurately represent the contents of their full-text counterparts. Participant 2 from the naturalistic study suggested that although some abstracts were “right on,” others either “overvalued” or “undervalued” the full-texts. It is pointed out in Chapter 2 and Chapter 6 that IR researchers should study ways to create a high proportion of “right on” abstracts and reduce the amount of “overvalued” and “undervalued” abstracts.

Because of the frequently questionable representational quality of abstracts, one alternative that I offer for consideration is to create an interactive system that allows users to instantly create document summaries based on their own needs. With this system, users would have control of what pieces of information they want in the document summary for a given full-text. Upon receiving input for the specific ingredients for the summary, the system would retrieve text segments that include specified elements and integrate them into a display. For instance, suppose a user wants to have a document summary that contains only the purpose, procedure and statistical analysis method. The system would retrieve paragraphs

that contain the terms “purpose,” “procedure” and “statistical analysis.” This approach is similar to the KWIC (Keyword in Context) format in many of the current full-text systems. However, the system that I envision would have coded the paragraphs of a full-text into semantic units. In this way, instead of getting arrays of text segments that usually have no logical continuation, users would obtain text units that describe the purpose, procedure and statistical analysis of the article. This design may work relatively easier with scientifically oriented articles such as medical literature since these texts contain structures that support more efficient decomposition or segmentation of texts. With the development of current web technology to support the metadata capability of SGML and XML in assigning DTD (document type definition) to the full-texts, this kind of text manipulation seems probable.

I also envision that this interactive document summary would be offered in addition to abstracts. In the case that users are not satisfied with the abstracts and do not want to read the full-texts, an alternative way for people to get what they want to know about the full-text would be to use this self-generated document summary to suit their needs.

Implications for Methodological Development

An important implication of this study is that multiple methods should be used for the investigation of people's criteria. This study used a laboratory experiment to measure the change in the use of criteria by having participants rate the importance of the 15 predefined criteria immediately after they read abstracts and full-texts. Meanwhile, a naturalistic study was conducted to investigate change in use of criteria by measuring the frequency of mention by the participants in their oral and written evaluations. The designs of the two studies differ in many respects from one another, and consequently, the results of the two studies are not directly comparable. However, both the commonalities and the differences between two sets of findings provide some insights to better understanding the notion of relevance, the process of document evaluation, and the actual use of criteria.

This study also derived a classification system that categorizes criteria by their nature: Objective and Subjective. Along with this classification, the researcher proposes that the appropriateness of a research method may also depend on what type of criteria it is to investigate. In other words, a laboratory and controlled design may be more appropriate to examine the use of Objective Criteria, whereas Subjective Criteria are better studied in a naturalistic environment with little control and manipulation to the research setting. Of course, the best way to study the use of criteria is not to mechanically separate the criteria, but to investigate how they are used together and interact with one another. An idealistic design would be combining multiple methods into one research setting, i.e., imposing certain

controls and structures in the process, in the meantime incorporating qualitative ways of collecting data.

Implication for Future Research

This research produced interesting findings regarding change in the use of relevance criteria across the two stages of document evaluation. The study found several new research questions that are worth pursuing. Below I will describe some of these research questions.

This study concluded with a new classification that categorizes criteria by their nature (Objective versus Subjective) and by their functionality (Primary and Secondary). The Process Model of Relevance Judgments was revised based on this taxonomy of criteria. The model now predicts a trend moving from an objective orientation at Stage 1 to a more combined view of objective and subjective at Stage 2. Subjective Criteria have a stronger tone for Stage 2 full-text evaluation.

The new taxonomy of criteria needs to be operationalized in empirical contexts to test its research practicality. With a series of empirical investigations, it is hoped that this taxonomy of criteria be further modified and developed based on the support of empirical evidence.

The second research question has to do with investigating *Topicality* as a class of criteria. This study found that *Topicality* is an important and frequently used criteria class for both the record evaluation and the full-text evaluation. It is argued that *Topicality* is a complex, multi-attribute factor, and the identification of the elements of this factor is difficult since *Topicality* is highly situational dependent. It

is also claimed that *Topicality* often interacts with other criteria. Despite of the complexity and subtlety involved with this factor, it would be useful to investigate the nature and properties of *Topicality*, particularly how the topicality varies for the two document representations. Because *Topicality* is the fundamental factor of relevance, research effort devoted to this factor would produce empirical findings that are useful to developing theories of relevance in particular and the process of information retrieval in general.

The third research question concerns “Interestingness.” As a frequently used criterion at both stages, “Interestingness” seems to denote different meanings depending on the context of use. “Interestingness” also seems to be distinguishable from “Topicality” or “Usefulness.” One potential research question is to investigate the association of the three criteria, i.e., “Interestingness,” “Topicality” and “Usefulness.”

A similar question may be posed to about “Newness,” or, in Barry’s term, “Content Novelty.” “Newness” was viewed as highly valuable during the document evaluation process for the naturalistic participants of this study. It would be interesting to examine whether “Newness” is viewed as important and used as frequently in other situations. In other words, the question is whether the ratings of importance and frequency of use for “Newness” varies by participants’ research experience and levels of education.

A final research question would be to explore the criteria that are similar to those grouped in this study under the class *My Study*. It seemed that the

components of this factor help to define topical relevance; they also reflect what constitute *Utility*. Furthermore, the examination of the elements in *My Study* could very well capture the movements or transformation in scholarly research (from others' research to "my" research). Finally, studying this specific class of criteria would provide insights into "relevance" and "usefulness."

APPENDIX A
UNC AA-IRB (Academic Affairs Institutional Review Board)
APPROVAL FOR LABORATORY EXPERIMENT

APPENDIX B
AA-IRB (Academic Affairs Institutional Review Board)
APPROVAL FOR NATURALISTIC STUDY

APPENDIX C
PRE-EVALUATION INFORMATION SHEET

Please tell us the following information about yourself:

Name:

Gender: Female Male

Class Level: Freshman Sophomore Junior
Senior
 Other (Please specify)

Major:

How much do you know about the "Year 2000 Problem"? (Check all that apply)

- Never heard of it
- Heard about it once or twice through reading newspapers, watching television, or browsing the Internet
- Heard about it a lot through reading newspapers, watching television, or browsing the Internet
- Have been very concerned about it
- Discussed it formally in class or informally with family members, friends, and colleagues
- Conducted research on it and wrote a report about it
- Other (please explain)

APPENDIX D
ABSTRACT CHECKLIST

Check the titles of all articles that you want to read before your prepare your outline.
(Note: you must select at least one article.)

Please also circle the titles of any articles you had already read before participating in this study.

- 01: Entering the Black Zone
- 02: Year 2000 conversion: been there, done that
- 03: Countdown to the millennium
- 04: The Y2K Crisis
- 05: Conversion crunch
- 06: Dear Mr. Gates
- 07: Once you get by all of the myths about the year 2000, you can focus on fixing it.
- 08: Government's time to act.
- 09: The hidden sides of Y2K
- 10: Imaging, bandwidth to drive technology plans
- 11: Industry wakes up to the year 2000 menace
- 12: How lethal is the millennium bug?
- 13: Is the year 2000 problem overhyped? Impossible!
- 14: What's all the Y2K panic about?
- 15: Year 2000 quirks will hit us slowly
- 16: Are you ready for year 2000?
- 17: Y2K: the scary part is not what you think
- 18: Year 2000: challenges, solutions & implication for the future Testing for 2000
- 19: Testing for 2000.
- 20: Year 2000: What? Me Worry?

APPENDIX E: ABSTRACT QUESTIONNAIRE

13. The documents seem to provide a clear definition of the Year 2000 Problem.

1	2	3	4	5	6	7
Not at all important						Extremely important

14. The documents provide information that is consistent with what I already know about the social effects of the Year 2000 Problem.

1	2	3	4	5	6	7
Not at all important						Extremely important

15. It seems likely that the information provided in these documents is accurate and trustworthy.

1	2	3	4	5	6	7
Not at all important						Extremely important

16. Other reasons. (Please describe any other reasons, and use the back of the sheet if needed. For each one list the number that you would circle on the scale of 1 to 7, where 7 is the most important.)

APPENDIX F
A SAMPLE PRESENTATION OUTLINE

APPENDIX G
A SAMPLE DOCUMENT EVALUATION SHEET

ARTICLE TITLE: Meta-analysis, clinical trials, and transferability of research results into practice. The case of cholesterol-lowering interventions in the secondary prevention of coronary heart disease.

1. After reading the article, please rate the article based on how useful it is to you. Please circle the appropriate number.

1	2	3	4	5	6	7
Not at all Extremely Useful						Useful

2. Please state the reasons that you believe that the article is as useful as you have described it.

APPENDIX H

SAMPLE DOCUMENT EVALUATION SHEETS COMPLETED BY PARTICIPANTS

Sample 1 (Participant 1).

ARTICLE TITLE: Origins and functions of positive and negative affect: A control-process view.

RATING: 6

REASONS: This review article was excellent! It beautifully extends their 1982 paper and accounts for why emotions emerge. The analysis helped hone my thinking about self-regulatory processes. They argue that there is a meta-monitoring feedback loop that examines the rate at which one narrows the discrepancy b/c goal attainment.

(Participant 5).

nd methodological issues in testing the circumplex structure of data in personality and social psychology

Rating: 6

-written and clear in its
of testing the

appropriateness of a circumplex model for a particular set of variables. They very

described how those assumptions translate into statistical assumptions. They went

various aspects of the fit of a circumplex model, including two examples of applying it to previously analyzed data. Although our data may not be sufficient to these procedures, this article provided a good conceptual foundation for evaluating whether our data conform to a circumplex model.

Sample 3

ARTICLE TITLE: The role of identifiability in the reduction of interindividual
tinuity.

REASONS:

This paper is very useful for the following reasons:

It involves a comparison between interindividual and intergroup interactions
-motive interdependence (as represented by the
ame)

It includes choice behavior (i.e., cooperation versus competition) as

c) It provides detailed statistical information regarding the difference in competitiveness between interindividual and intergroup relations, allowing for the computation of effect-
-analytic study.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

(Examples listed by PN [Participant Case Number])

Criteria Class 1: **Topicality**

Domain

Definition: The document is on the right/different domain.

Examples: P1: This is apparently they've got into a new domain. This is also not related.
P4: This is relevant, I can tell from the title. Because antecedents of distribute and procedural justice are exact the domain we are interested in. It's about procedural justice stuff.
P4: This would be interesting, it talks about procedural justice and in an arbitration fact-finding domain.

Is About

Definition: The participant describes what the document is about.

Examples: P4: This is more about the ethics of science in general. That's not relevant to what I am interested in.
P4: It's talking about the makeup of the groups and how that affects all the perceived fairness of the outcome. That would be relevant.
P5: I don't think this is too relevant, cause it's not really about first impressions, but let me just see here, it is about the five-factor model.
P5: No, it's not really relevant, cause it's about how the personality are related to their kinds of interactions or number of interaction.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 1: **Topicality** (Cont.)

Interest of the Study

Definition: The participant states the research interest of the study.

Examples: P5: ...Personality measurements of both the students and cooperating teachers were based on a measure unrelated to the Big Five (the Myers-Briggs Type Indicator) and were used primarily to categorize pairs of student and cooperating teachers according to their personality preferences. Attitudes toward teaching were also examined, which are of no interest to us.

P5: The reason that it wasn't really relevant is because that their main interest was in having people do a Q-sort to describe themselves, but to describe themselves in two ways, their actual self, how they really are, and their ideal self, how they would like to be. And these researchers were more interested in the correlations, the kind of agreement between those two sets of ratings, and we are not really interested in the ratings actual and ideal selves, we are interested in ratings of other people.

P5: ...these researchers' main interest was in sort of how people choose to get more information or to cumulate information about people. And they were trying to relate it to this particular characteristics of dogmatism, which is supposed to determine whether people would want to get more information or not. And we were not really interested in that.

Topical Focus

Definition: The participant comments on the specific focus of the study.

Examples: P3: I'd say no. This is focus more on application at academic setting.
P3: No. This is applied stuff, and it's focusing more on looking at the culture of a training program instead of the culture of the family.

Topical Relatedness

Definition: The study is generally related/unrelated to the participant's topic.

Examples: P1: It doesn't seem to be related to self-regulation. This is a personality variable.
P1: It dealt indirectly with the issue of self-control, so primarily it wasn't relevant to my topic in that it deals with a lot of other personality traits.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Research Structure

Argument

Examples: P5: This article is sort of arguing that Wiggin's circumplex model or the r the date that horizons from studies of that model could be explained by something a lot simpler that has basically to do with descriptive statements.

P5: The authors make some important points about the potential analyses. They argue that one popular circumplex model of personality to the socia

Concepts

Examples: P5: This was highly relevant as it centered on the underlying dural justice.

P3: This article also includes empirical data on reciprocal effects. terms of a coping response.

Conclusion

Definition: The participant examines the conclusion of the study.

Examples: P8: It was focused on interventions, but it was trying to cover every type of interventions that's ever been studied with women. And the sically was OK-- haven't tried any one thing with women with this problem enough to really evaluate it very well. Which was interesting to me because I ole lot done with psychosocial interventions for women. I was surprised to see area as well.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 2: **Research Structure** (Cont.)

Data

Definition: The participant sees whether the document contains data and the nature of the data.

Examples: P7: Gave me factual information on meds; didn't know before.
Provided me with combined effect sizes for cognitive therapy versus behavior therapy, which alleviates the problems with several of these studies in isolation – i.e., power.
P6: Well, I am going to mark that, but it's more a review paper than a paper that contains original data.
P4: I gave it a rating of 4 in that it's somewhat relevant, but it's kind of review paper or armchair paper in that not significant amount of data was provided to support their arguments.
P3: This article also includes empirical data on reciprocal effects. However, the subjects are adults & social support is not discussed in terms of a coping response.
P3: This article only includes theoretical arguments & no empirical data.

Design

Definition: The participant emphasizes the research design of the study.

Examples: P7: poor design, very superficial description of cases and procedures. Nothing new.
P7: Useful example of design. It's got no treatment, didn't really work very well. It's more behavioral treatment instead of...

Evidence of Effect

Definition: The participant determines whether the study provides evidence of effect, and whether the effects are significant .

Examples: P3: We got stress, and we got possible coping. I would say no here actually because they are not finding the reciprocal effects.
P8: So we do have an empirical situation where they offer an intervention and even though the effects aren't significant, they are viewing that these interventions do give some additional benefit over the usual medical care, so that looks interesting.

Implication

Definition: The participant states the implication of the study.

Examples: P2: Well-written -tackles many issues; more difficult to use in a chemical study.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS

Research Structure (Cont.)

Interpretation

Examples: P6: The usefulness of this paper may be limited by the fact that ...these studies involved variations of the traditional Prisoner's Dilemma Game
author(s) of the study.
cooperative and competitive choice is altered.

Method

Definition: The participant comments on the methodological aspects of the study.
Examples: P7: This is one of the earliest empirical examination of my research question ... It's flawed methodologica
useful results and suggested some interesting reasons why they found those results which helps me formulate more hypotheses for my own study

Nature of the Study

Definition: The participant describes the research
Examples: P3: This is a longitudinal study, and there aren't a lot of those out there.
P5: This is again an application of the circumplex to a particular, developing a particular measure or something like that.
P8: This is just a survey type of deal, doesn't seem to be very

Population

Examples: P3: This was Asian people, people from Kwan, I think, and they were , so wasn't the same.
P5: Well, this is almost relevant if it weren't focus on clinical psychiatric people. They are rating people who are very extreme and have some

Practicality

Examples: P8: It was an discreditably intense intervention, I mean it was very ffective, but I couldn't see how it would be very practical to implement.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 2: **Research Structure** (Cont.)

Procedure

Definition: The participant describes the procedure of the study.

Examples: P2: Actually this would be interesting because they are giving a local anesthetic 3 min before or 5 min after the formalin injection, now this is important because it's directly maps on the two studies I just talked about.

P5: Although much about the research design and statistical analyses in this article is of poor quality for their largely experimental purposes, the study was in many ways similar to our more exploratory, observational study. Students viewed short film clips of one other person and provided open-ended verbal responses explaining why they would or would not like to get to know the viewed person.

Research Assumption

Definition: The participant attends to research assumptions of the study.

Examples: P5: They very clearly laid out the substantive assumptions underlying circumplex models and described how those assumptions translate into statistical assumptions.

P5: So there are some assumptions about what the patterns of correlation should be among a bunch of variables, there are some assumptions what need to be meant for something to be a circumplex conceptually, and then they made it very clear that those assumptions, the conceptual assumptions imply, lead to some fairly specific statistical assumptions. Like the statistical characteristics that could be tested.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 2: **Research Structure** (Cont.)

Results

Definition: The participant discusses the findings of the study.

Examples: P3: What they found was there was evidence for reciprocal effect, which I was arguing, that there was probably the case in my study, and they found that emotional restraint, which is a coping mechanism, did seem to decrease the likelihood for association with drug using peers or mental adjustment, which is what I found too, that certain avoidance was better for externalizing problems.

P3: The only thing that was really of interest to me from this article was that they found that frequency of the exposure to whatever the stress that you are studying have very important implications for the coping strategies that's used.

P4: This article also includes empirical data on reciprocal effects. However, the subjects are adults & social support is not discussed in terms of a coping response. The authors did find, however, that environmental effects (in my case, stress) seem to exert greater influence on mental processes (in my case, internalizing & externalizing problem behavior) than vice versa.

P7: This is one of the earliest empirical examination of my research question ... It's flawed methodologically, but that's ok ... it found some useful results and suggested some interesting reasons why they found those results – which helps me formulate more hypotheses for my own study.

Sample Size

Definition: The participant attends to the aspect of the sample size in the study.

Examples: P8: This is good to start with, because we are dealing with population of women who already have been diagnosed with heart disease. Got a large sample.

P8: The bad side of it, was there were only 15 women or so out of well over 100 The participant in this study.

State of Research

Definition: The study refers to the current state of research.

Examples: P1: I think people used the term interchangeably... it kind of gives me a broader sense of what's out there.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 2: **Research Structure** (Cont.)

Statistic Analysis

Definition: The participant focuses on the statistical analysis aspect of the study.

Examples: P1: They have incredibly complex correlation and mediational analyses...
P3: The data analysis was quite unsophisticated.
P5: Principle Component Analysis has some undesirable properties or behaves, it does some undesirable things, and I don't understand why they use it... I think they have been using inappropriate path/factor analysis.

Techniques

Definition: The participant describes the techniques that were applied in the study.

Examples: P2: This is using preemptive analgesia for a surgical incision of laparoscopy technique...I am assuming it is some sort of biopsy technique or something like that. So this is definitely an article that I would want because it is actually using a local anesthetic technique that's administrated before skin incision.

Theoretical Model

Definition: The participant discusses the theories or conceptual model applied in the study.

Examples: P5: On one hand, the particular circumplex being evaluated in this study was Wiggins's Circumplex, which differs substantially from the AB5C circumplex used in our research. However, it did provide some insight into how one goes about testing whether the relationships among a set of variables conforms to a circumplex model.

Treatment

Definition: The participant describes the treatment used.

Examples: P7: ...mixes up behavioral treatment & cognitive treatment, meaning, gives treatment to subjects with both components at once, which is disappointing because I'm interested in cognitive treatment.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 2: Research Structure (Cont.)

Variables and Constructs

Definition: The participant focuses on the variables and constructs of the study.

Examples: P3: But they are also using a lot of the same variables that we are interested in, the social addiction...

P5: The construct is anger, not very relevant. Looks like they do try to relate to personality. But...they are focusing on a different construct.

P5: The construct here is mood, two different ways of measuring it, and they are associating it with multidimensional scaling. Not relevant. Cause its construct is too different.

Criteria Class 3: Quality of Information

Accuracy

Definition: The participant's evaluative statement regarding whether the information is reflected accurately or the methods applied are executed accurately.

Examples: P1: This article seems to imply that self-regulation isn't a depleteable resource... I just didn't have all that faith that it was done accurately.

Background Information

Definition: The item provide some background information on the topic.

Examples: P2: Early history of issues of pre-emptive analgesia where everything started - excellent background.

P2: So this was very nice, just like a background article.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 3: **Quality of Information** (Cont.)

Clarity and Well-Written

Definition: The participant's evaluative statements regarding the overall quality of the paper from the view points of clarity and writing style.

Examples: P5: This methodological article was very well-written and clear in its explication of the disadvantages faced by conventional methods of testing the appropriateness of a circumplex model for a particular set of variables.

P5: I don't know if it is because they are Dutch, and just there are some language differences in their writing style, but they were using a lot of terminology that seem to be very loose kind of terminology and they didn't define things very well, they used terms that sort of roughly seem to I could understand, but it could've two or three things, what they were saying could've meant, I mean the implications of what they are saying would've been very different if they meant this thing versus this other thing.

Completeness

Definition: The participant's evaluative statements regarding whether the information provided in completed or not.

Examples: P3: And it's because it doesn't have all the pieces of the model there, it has the religiosity, which is a coping piece, and delinquency, which is the outcome or the adjustment piece, but it doesn't have the environmental stuff: the stress or the peer group thing.

P5: This was a literally short article...it's just in an abstract form, really, and there is not that much detail in it...I gave that a rating of 3, just because largely because there just wasn't enough detail to know what exactly they did, it was a really condensed form of the study.

Didn't read

Definition: The participant states that the item was not read due to the seemingly poor quality of the information presented.

Examples: P8: This was a response to the prior article, and I rated as a 1, because I didn't even read it, because I had read the first one and it was of no good to me, so I just kind of ignored this one.

P8: I didn't even read it, I mean I glanced through it, but I didn't go through it very clearly, cause I could tell by glance through that it wasn't going to contribute much to it.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 3: **Quality of Information** (Cont.)

Importance

Definition: The participant's evaluative statement regarding the importance of the study or the importance of the information presented in the document.

Examples: P1: It seems to be important. Because it seems to me that this is going to really summarize their perspective that control, both failure of self-control and failure of self-regulation, I think people used the term interchangeably, is the cause of aggression.
P1: Why did I think this one was important? It was theoretically very interesting and methodological very intriguing, and perhaps these are methods that I can actually use. There are complicated and intricate, but brilliant, they are methods that I actually be able to use.
P5: So this is definitely important to read, because it's the first place where that term comes up...it's important for me to understand the zero acquaintance situation first.

Insightfulness

Definition: The participant's evaluative statements regarding the insightfulness of the information presented in the document.

Examples: P5: It's just sort of a little bit of insight into what the big five might be related to.

Level

Definition: The participant's evaluative statements regarding the level of the information presented in the document.

Examples: P7: No, too elementary.
P7: No, too introductory.

Quality and Value

Definition: The participant's evaluative comments on the quality of the information presented or overall value of the study.

Examples: P1: This review article was excellent! It beautifully extends their 1982 paper and accounts for why [emotions] emerge. The analysis helped hone my thinking about self-regulatory processes. They argue that there is a meta-monitoring feedback loop that examines the rate at which one narrows the discrepancy b/c goal attainment.
P1: Probably the most brilliant earlier work on this stuff.
P3: This book chapter does an excellent job of summarizing much of the literature that I have reviewed for my thesis.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 3: **Quality of Information** (Cont.)

Readability

Definition: The participant's evaluative statements regarding the readability of the document.

Examples: P2: And also I guess they are doing some types of It's really hard to follow, the language is really hard to follow...
P2: Excellently controlled study. And pretty easy to read, too.

Repeat

Definition: The participant claims that the content of document is repetitive or redundant.

Examples: P1: This article was extremely good, but it largely repeated the precious article. Hence, it was not especially useful to me.
P8: I didn't rated this as more useful, simply because the effect it did give me, I'd already gotten elsewhere, I mean we already kind of knew this. So it's a little bit redundant.

Scope

Definition: The participant's evaluative comments on the scope or coverage of the document.

Examples: P4: I can tell from the title probably it's too board, but ... yeah, this is more about the ethics of science in general. That's not relevant to what I am interested in.
P5: But it looks like it's focusing on pretty specific on agreement between your own rating of yourself and other's ratings of yourself. And so, again, it's a little bit too much on consensus, so I don't think I really want that.
P5: It's certainly interesting, I mean, it's interesting to me but it's talking about a somewhat different circumplex, I mean, it's talking about a very specific circumplex model.

Sophistication

Definition: The participant's evaluative statements regarding the depth of the information provided.

Examples: P7: Poor design, very superficial description of cases and procedures. Nothing new.
P7: Describes a very different type of treatment. Discussion of cognition very minimal.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 3: **Quality of Information** (Cont.)

Specificity

Definition: The participant's evaluative comments on the concreteness or specificity of the information presented in the document.

Examples: P1: This article seemed too specific and applied for my purposes. It dealt with setting goals for productivity in the workplace. Still, it does suggest that perhaps self-regulation doesn't interfere with later performance.
P2: This article is a more general article on postoperative pain management. And they are talking about the important role of preemptive analgesia. This is a particular relevant article for citing, it is not going to have a lot of data that is going to be of particular interest, but probably going to give me a lot of backward references I would imagine, back citations.

Starting Point

Definition: The document provides a starting point for The participant' project.

Examples: P2: So this would be a really, really good, actually probably an excellent starting point because it is a 1996 review so it is relatively recent.

Strangeness

Definition: The participant's evaluative statements regarding the normality of the information presented in the document.

Examples: P5: Something about note-taking behavior. Yeah, it's wired, it's not relevant.
P5: But his data collection methods, like how he justify and how he was collecting data and particularly his statistical procedure, he was doing a lot of wired, obscured statistical tests, and it wasn't obvious they were testing what he wanted to test, and he didn't really it really well, so I am not really convinced that anything he was finding is statistically significant or meaningful.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 3: **Quality of Information** (Cont.)

Suitable for Meta-analysis

Definition: The participant's evaluative statements regarding whether the data or information contained in the document is appropriate for a meta-analysis project.

Examples: P6: These particular variations of the PDG are used only in this study, which makes them unlikely candidates for the role of mediator in a meta-analytic review. Again, these studies are somewhat unique and can not easily be categorized into a category containing multiple studies, which would make it useful for a meta-analysis.
P4: So that itself is useful, but the book review can not be entered in meta-analysis. Actually it was probably my error in selecting that initially.
P4: I gave it a rating of 4 in that it's somewhat relevant, but it's kind of review paper or armchair paper in that not significant amount of data was provided to support their arguments. On the surf of purpose of the meta-analysis that could be a problem.

Title Indicativeness

Definition: The participant points out whether the title is indicative to reflect the relevance of the actual document.

Examples: P4: This is relevant, I can tell from the title. Because antecedents of distribute and procedural justice are exact the domain we are interested in.
P9: I knew that's going to be a good one. So I am going to click on that. I don't need to read the abstract.
P9: Just by the title it doesn't seem like there is anything there that would tie me to look at the abstract.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 3: **Quality of Information** (Cont.)

Trustworthiness

Definition: The participant's evaluative statement regarding whether the empirical manipulation reported is trustworthy or believable.

Examples: P1: I just don't have all that much faith that this was a good manipulation.
P5: They were very clear about what exactly is being tested, what their terminology is, and what it does and does not mean. It just made it very believable that you would be testing something meaningful if you follow their approach.
P5: ...well if this is how people are justifying using a circumplex model, I don't think I believe it, I don't think I'd buy it.

Uniqueness

Definition: The participant claims that the study reported is unique.

Examples: P2: So this is a really interesting study and again they did a really good design here because they did preincision and postincision, and the problem was they used preincision large dose and postincision moderate dose, but still it is a very, very interesting and unique study.

Criteria Class 4: **Source Value**

Article Type

Definition: The participant defines the document type.

Examples: P5: It's a book review. The problem here is that there is no data. They are just commenting on someone's book.
P3: Lot of references in this, this is a review, basically. Um...the only drawback from this is that it really involved health psychology, so it wasn't psychological sorts of things, like internalizing problems externalizing, and it's all adults too.
P2: This is a literature review on the concept of "preemptive analgesia."

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 4: **Source Value** (Cont.)

Author

Definition: The participant mentions the author(s) of the document.

Examples: P1: This is by Higgins and Kruglanski. They are very good.
P4: This again was coauthored by Tom Tyler. It's about trust with authorities and people under them, that's definitely relevant.
P5: This one I want, but not for related to this. I think it just gives some methodology and statistical procedures for actually getting measures of agreement or consensus in this type of research. And Pet SHROUT is just well known in that area. That's the only reason to I'm keeping that.

Author Bias

Definition: The participant's evaluative comments on the political biases exhibited by the author(s) of the document.

Examples: P8: And this I didn't give it a 7, just because the author had some biases that were so powerful that I wasn't sure how much I trusted her totally. I mean I consider myself a very strong feminist, but she was going over the line in a couple of the ways that I get nervous with people who are very...She had this thing about women due to their inherent subordinate status in their patriotic society and so forth.

Cited Author

Definition: The participant comments on the authors/researchers cited in the document.

Examples: P5: Yeah, I don't think it's relevant. The only reason that that came up is they must have cited Breckler, et al. for using multidimensional scaling cause they are using multidimensional scaling here also.
P5: They give Hogan in it, 1983 as a reference for "the Interpersonal Circumplex." I don't know it that's...if he came up with this interpersonal circumplex model or a particular way of measuring it, or what. But I've heard of Hogan. That seems too specific.

Cited Frequently

Definition: The participant comments on the citing frequency of the document.

Examples: P9: Wow, this is cited 181 times. Yeah, this looks kind of interesting, by identifying negotiation.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 4: **Source Value** (Cont.)

Cited in Preliminary Paper

Definition: The document was cited in participant's preliminary paper.

Examples: P5: So this is the one that we cited in the article. I suppose I should get this, because we cited it, so I need to know what's in here versus in other things we might want. Ah...consensus, what is this consensus? Oh see, this particular article is about how different people agree or not on their ratings of someone they've never met...this is the article we cited, Kenny, et al...

Classic Study

Definition: The document being reviewed was the first article on the topic or one of the classic studies on the topic.

Examples: P7: This is one of the earliest empirical examination of my research question ... It's flawed methodologically, but that's ok ... it found some useful results and suggested some interesting reasons why they found those results...
P5: This might be the place where that terminology was coined...OK, this is really coin it... So this is definitely important to read, because it's the first place where that term comes up...

Geographic Location

Definition: The participant mentions on the geographic location of the study.

Examples: P8: This is like so old, 1972-1976. In the West of Scotland, I didn't know there were that many people in Scotland. OK, this is just a survey type of deal, doesn't seem to be very empirically oriented.

Item Mentioned by Coauthor

Definition: The document was mentioned by the coauthor as valuable for revising participant's preliminary paper.

Examples: P5: OK. I am almost positive that this is going to be relevant, because what it says, because I am supposed to read this as for one of my coauthors about the "thin slices" stuff. And it does look relevant.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 4: **Source Value** (Cont.)

Journal

Definition: The participant make remarks on the journal that the document was published in.

Examples: P3: I would say no here actually because they are not finding the reciprocal effects. And not in a journal that I heard of before.
P1: This looks like it could be relevant, but it's in the journal named "Anxiety,-Stress-and-Coping:-An-International-Journal." And it doesn't look like that new or to add too much to my personal memory, so I am just going to skip it.

Language

Definition: The participant comments on the language document was written in.

Examples: P5: It's a multidimensional scaling, and it is personality. This is an article in Spanish, I think. And it doesn't look terribly relevant, so I am going to skip that.
P2: I might want to skip number nine just because it in another language.
P2: I actually read German. So it would be OK. And the reason that this would a very important study for me because this is another drug.

Length of Article

Definition: The participant mentions the length of the document.

Examples: P8: They are both one to two page articles. These might be interesting for me to read it in terms of again introduction sorts of material and providing a basis for why I am doing this, but I doubt that it will give me that much high way of actual information.
P5: This was a literally short article, it turned out that it was a part of a proceedings for upcoming convention, so it's just in an abstract form, really, and there is not that much detail in it.

Only Title Available

Definition: Bibliographic records that list the title of the document but not the abstract.

Examples: P4: I don't think this is relevant just judging from the title, but there is also no abstract, so I can't make any inferences further.
P4: Yes, it's relevant. Because it's about procedural justice. There is only title that I can go by.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 4: **Source Value** (Cont.)

Publication Date

Definition: The participant notes the publication date of the document.

Examples: P8: This is like so old, 1972-1976. In the West of Scotland...
P5: Note this was published only two-year after...1992 was the original AB5C Hofsee, De-raad, and Goldberg publication. So this is only two years after that.
P5: Although it's back in 79, looks like they generated some simulated data from people personality response.

Reference

Definition: The participant comments on whether the document lead to useful references to other documents.

Examples: P5: The two things I am interested in, some kind of circumplex and the Big-5, so for now that seems relevant. It might have some good references in there, which will be good.
P5: This study was somewhat relevant in that it reviewed a large number of studies that had examined student ratings of teacher personality; this provides some useful references for further reading on this issue.
P8: I will look this up in Social Science Citation Index, mostly to get its citation, because I doubt it will be cited in much of any place yet. Cause it's so recent. But it should have some very helpful reference for me to look at, since this is reviewing all of these studies.

Referenced in Items Selected

Definition: The document was referenced in a document selected previously.

Examples: P5: And it's borderline relevant, but I think it's too specific, and it will be referenced in later articles I already picked I think so I can go back to it if I really want to.

See Items Cited This

Definition: The participant wants to see what documents cited the document being reviewed.

Examples: P9: Just got a title. And I will go to check on that one, and I want to see who cited it.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 5: **Cognitive State**

Add My Knowledge

- Definition:* The participant acknowledges that the document adds something to their knowledge base.
- Examples:* P1: This looks like it could be relevant, but it's in the journal named "Anxiety,-Stress-and-Coping:-An-International-Journal." And it doesn't look like that new or to add too much to my personal memory, so I am just going to skip it.
P9: Although this article doesn't add much to lit, it would be useful for a review.

Certainty

- Definition:* The participant expresses the degree of certainty in decision making.
- Examples:* P1: That could be relevant...I am not sure if this is relevant yet. See, I don't know how they used self-regulation, I don't know, I can't quite tell.
P1: I am not sure it's relevant or not. It's a book chapter. I might want to skip, We have a lot of book chapters, this is kind of conjectural. Could be relevant, I am not sure.

Expectation

- Definition:* The participant describes the initial expectation of the document when they first read the bibliographic record of the document.
- Examples:* P3: Umm...This one, if I remember correctly, I was hoping that I would get some information about the way that families help their children to cope with stressful problems, and...it wasn't really focused on that...
P3: Again, this is a 1. I was hoping that there'll be some discussion of at least the way that black parents socialize their kids to cope, but there wasn't.

Familiarity

- Definition:* The participant indicates how familiar they are with the document or the theories presented in the document.
- Examples:* P1: That definitely looks familiar. This is good.
P1: Oh, I never heard of this at all. I could be less ignorant on this...

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 5: **Cognitive State** (Cont.)

Informativeness

Definition: The participant evaluates the documents based on the amount of the useful information contained, and hence how informative the document is.

Examples: P2: The first article is going to be much more of a informatory, probably theoretical model.

P2: So it was not the one of the best as far as information...

P5: Another reason this study may be informative for ours is that a wide variety of purely physical nonverbal behaviors (e.g., laugh, fidget with hands) were extracted from the videotapes by coders other than the subjects...

P7: This was like very, very basic, elementary...even the description was very superficial. So no information at all basically.

Inspirational

Definition: The participant indicates how much the document has inspired them and honed their thinking.

Examples: P1: There are articles that are useful for the project at hand because they hone my thinking. Even though they are not totally related, they make me think clearer about the issue of self-regulation. So this still would be in the 5-6 range, even on that criterion.

P3: Although this sample was vastly different from my own...This relates to the conclusions made in my study, but also points to some new directions for analyzing my data (i.e., look at other cognitive moderators/mediators).

P5: It was a very provocative article, cause it made me ask a lot of questions about well if this is how people are justifying using a circumplex model, I don't think I believe it, I don't think I'd buy it.

P7: This article is a theoretical paper, and it's very useful in arguing why my theory is flawed. It's very thought provoking and help me get a better grasp on why I believe what I believe by providing the opposite opinion.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 5: **Cognitive State** (Cont.)

Interestingness

Definition: The participant states that the document is interesting/uninteresting.

Examples: P7: No, off topic although it looks interesting.
P8: I rated this as a 5, because it did do something very interesting, it compared two types of interventions, one was a more psychosocial stress management sort of intervention, and it compared that to an exercise program.
P9: OK, this looks interesting, here people are making a choice, and they say they measure self-esteem, but they don't really show how self-esteem is predicted...
P5: It's not the five factor model, and it's more of a computer simulation than it's really a theoretical explanation of what the circumplex is. Interesting, but not too relevant for this.
P4: This would be interesting in terms of looking at procedural justice from this human right angle.
P2: This is one of the ones that I gave a 7 to. I gave this one a 7 because they have very, very interesting results here...
P3: Even though it is a different population, I think it is interesting, because what we have is a different population too.
P1: And it has some interesting and somewhat relevant ideas but this article focused primarily focused on a different issue...and it seems to happen at a subconscious or unconscious level, which is interesting.

Read Before

Definition: The participant indicates that he/she has already read the document or had the document.

Examples: P9: This one I read a long time ago, but I know it's self-enhancement in romantic relationships rather than self-verification, so I am going to click on that.
P5: This is the original article I keep referring to. And I already have this, it's very relevant. But I already have it.
P1: This is one of the articles that I have, looks good.

Remembering

Definition: The participant indicates how well he/she remember the document.

Examples: P2: I don't remember this article very well. I think this was another morphine article. The design wasn't as good if I'm remembering it right...It just wasn't as helpful as some of the others.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 5: **Cognitive State** (Cont.)

Understandability

Definition: The participant indicates being able or unable to understand the document.

Examples: P1: I don't really understand this stuff. That's not related to self-regulation either.

P5: That might be interesting, but it's too far beyond what I understand right now. I don't think it will be too useful

P5: This looks like that it's making theoretical extensions of that or some kind of extensions of that, but at this point for me it's too far beyond what I am trying to understand about the "zero acquaintance"...

Agreeability

Definition: The participant agrees with a particular set of propositions in the document.

Examples: P1: Let me just make sure. "The resource allocation model of goal setting maintains that self-regulation initiated through goal setting requires attentional resources that could be more productively applied to skill acquisition and complex task performance." I don't know if I agree with that.

Newness

Definition: The participant indicates whether the information content of the document is new or not new to him/her.

Examples: P1: The other reason that why I didn't give it a higher value, is I knew this stuff before. So I thought I was rereading this that wasn't as new to me.

P1: Overall, however, I gave the paper a "4," and again this is kind of my "newness" criterion, I knew all this stuff...

P7: Poor design, very superficial description of cases and procedures. Nothing new.

P7: I gave it a 1. Which is a comment that had nothing new, they are arguing about the points that are really not even important parts of the theory.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 5: **Cognitive State** (Cont.)

Originality

Definition: The document contains the original conceptualization of a theory.

Examples: P2: I rated it as a 7, it is extremely useful because the ideas of preemptive analgesia have been basically associated with a man named Patrick Wall in 1980. This guy was at 1913, and he did it first. And I did not know that... He said what he called was anoci-association because what we call pain fiber or nerve acceptor, they are carrying nausea information...

Support My View

Definition: The document provides support to participant's research framework and design.

Examples: P3: So they found that there was reciprocal effect, which is similar to what I am arguing in my study...

P7: The article addresses one component of the treatment that I am doing in my study, and it has experimental evidence to show why my manipulation might work, and it had a lot of population. So it's useful for that reason.

Criteria Class 6:

Affective

Definition: The participant expresses his/her affective or emotional reaction to

Examples: P2: This is a big deal, very nice study, like that very much.

P2: This is one of my favorites.

partners I think I have been a little too biased because I saw self verify another line of research.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Utility

Helpfulness

Definition: The participant indicates whether the helpful to him/her.

patients received preemptive analgesia. So there is no comparison here. It's not going to be helpful. Nice idea in won't help me at all.

what I like about this is that they used three different methods, factor analysis, cluster analysis, and multidimensional scaling, which we

ratings of other people.

The participant indicates whether the document is useful or not useful to him/her.

Examples: P5: This might be useful, because it will talk some about, I guess, the

P5: I don't think it's relevant. I mean it is probably relevant, but it's not going to add too much, not useful.

P2: This one was very, very useful for one reason. They did what a

radical and unique design...this was one of the best studies because it

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 8: **My Study**

Influenced My Study

Definition: The participant indicates that the document influence his/her own research via changing some elements in the research.

Examples: P7: Suggests evidence for including another specific measure in my study, also provides another way of calculating change that makes a lot of sense. Also have good references...I think I would give some thing a 7 if it actually change what I was doing in my study, influence me in that way.

P7: I gave it a 7. the critical article ...1st well-controlled study that shows that treatment I'm interested may be better than current treatment of choice. Lots of questions in discussion that my study can address. Lots of methods of assessment that I want to use now in my study.

P7: It found some useful results and suggested some interesting reasons why they found those result - which helps me formulate more hypotheses for my own study

Is What I Want

Definition: The participant indicate that the document is exactly what he/she want or doesn't want.

Examples: P9: This looks exactly what I want, just based on the title.
P9: Although it had a good title, which is why I and perhaps a good abstract but when I went through this article, it didn't really have what I was looking for.

Justification of My Study

Definition: The participant indicates that the document provide a good justification or rational for his/her research projects.

Examples: P8: Just as a justification for why I would want to do this study. Because from the abstract it sounds like that there indeed differences between men and women.

P8: So it was an useful article in my review paper my thesis paper because I can use it to make a point that what I am talking about is important that it is something that needs to be addressed.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 8: **My Study** (Cont.)

Link to My Study

Definition: The participant suggests the connection between the study reported in the document being reviewed and his/her own research projects.

Examples: P2: So this is a really, really interesting article and this is exactly the kind of study that I would want to do in a much more controlled environment than I would ever be able to use.

P3: Although the authors investigate the reciprocal effects of academic achievement & self-concept, there's no link to my research...

P3: Although this sample was vastly different from my own, the article is relevant to my research because it supports the notion that certain cognitive/behavioral factors (in this case, general self-efficiency) moderate the association between stress - adjustment. This relates to the conclusions made in my study, but also points to some new directions for analyzing my data (i.e., look at other cognitive moderators/mediators).

P7: The article addresses one component of the treatment that I am doing in my study, and it has experimental evidence to show why my manipulation might work, and it had a lot of population. So it's useful for that reason.

Personal Interest

Definition: The document being is not necessarily relevant to the topic, but the participant has own interest in it.

Examples: P5: This was the one you were just pick up for me, it was not really relevant to the study, but you just got this for my own interest.

P9: Although this article contained ego-confirmation stuff, it focused on process rather than looking at motives underlying interaction choices. This article has some research in there that was relevant to my own, but it looked too much on behavioral confirmation processes rather than anything else.

APPENDIX I: CODING SCHEME, DEFINITIONS AND EXAMPLES
OF THE CRITERIA MENTIONED BY PARTICIPANTS
IN THE NATURALISTIC STUDY

Criteria Class 8: **My Study** (Cont.)

Similar to What I Do

Definition: The participant makes comments on the similarity between the study reviewed and his/her own research project.

Examples: P3: This is an example of something similar to what I am doing, which is taking a different population, different ethnic group, and looking at the variables...
P8: This is very close to the interventions we did in terms of what they are interested in. So that's really cool.
P9: very useful. Same design as my study but my technique is more inclusive and broad.
P5: The most relevant thing was that they were looking at students' ratings of teachers, like us, in fairly limited acquaintance situations, like us. Small differences, they were rating, these were students who had been in the class for a while, and they were rating a guest lecturer who came in for a guest lecture. As opposed to rating the teacher they know is going to be their teacher on the first day of class.

Would Cite

Definition: The participant indicates that he/she plans to cite the document being reviewed.

Examples: P1: It's something that I would cite, when I wrote paper. I would put in my introduction section and say that some authors have noted that using self-regulatory strength to set goals doesn't interfere with later performance.
P9: This has to do self-verification, how people try to change people's Views so that's consistent with their own. So I am going to cite that.

Reading

Definition: The participant describes the reading experience.

Examples: P1: And I read every single word and highlighted extensively from a 20 page American Psychologist's article. So this took me many, many hours to read.
P1: This was the 7th one that I read. I gave this article a 6.
P1: The last two were actually the first two articles that I read. They probably were two the most exciting one.

REFERENCES

- Barry, C. L. (1993). *The identification of user criteria of relevance and document characteristics: Beyond the topical approach to information retrieval*. Doctoral dissertation, Syracuse University.
- Barry, C. L. (1994). User-defined relevance criteria: An exploratory study. *Journal of the American Society for Information Science*, 45, 149-159.
- Barry, C. L. (1998). Document representations and clues to document relevance. *Journal of the American Society for Information Science*, 49(14), 1293-1303.
- Barry, C. L., & Schamber, L. (1998). User's criteria for relevance evaluation: A cross-situational comparison. *Information Processing and Management*, 34(2/3), 219-236.
- Bateman, J. (1998a). Changes in relevance criteria: A longitudinal study. *Proceedings of the 61st Annual Meeting of the American Society for Information Science*, 35, 23-32.
- Bateman, J. (1998b). *Modeling changes in end-user relevance criteria: An information seeking study*. Doctoral dissertation, University of North Texas.
- Beaugrande, R. D. (1980). *Text, discourse, and process*. Norwood, NJ: Ablex.
- Beaugrande, R. D., & Dressler, W. U. (1981). *Introduction to text linguistics*. London: Longman Group Limited.
- Beghol, C. (1986). Bibliographic classification theory and text linguistics: Aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation*, 42(2), 84-113.
- Billings, R. S., & Scherer, L. L. (1988). The effects of response mode and importance on decision-making strategies: Judgment versus choice. *Organizational Behavior and Human Decision*, 41, 1-19.
- Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communication of the ACM*, 28(3), 289-299.
- Blair, D. C. (1996). STAIRS Redux: Thoughts on the STAIRS evaluation, ten years after. *Journal of the American Society for Information Science*, 47(1), 4-22.

- Boyce, B. (1982). Beyond topicality: A two stage view of relevance and the retrieval process. *Information Processing and Management*. 18(3), 105-109.
- Bruner, J. S. (1973). *Beyond the information given: Studies in the psychology of knowing*. New York: W. W. Norton & Company.
- Cattell, B. R. (1978). *The Scientific use of factor analysis in behavioral and life sciences*. New York: Plenum Press.
- Comrey, A. L. (1973). *A first course in factor analysis*. New York: Academic Press.
- Cool, C., Belkin, N. J., & Kant, P. B. (1993). Characteristics of texts affecting relevance judgments. IN M. E. Williams (Ed.), *Proceedings of the 14th National Online meeting* (pp. 77-84). Medford, NJ: Learned Information, Inc.
- Cooper, W. S. (1971). A definition of relevance for information retrieval. *Information Storage and Retrieval*. 7(1), 19-37.
- Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness. *Journal of the American Society for Information Science*. 24(2), 87-100.
- Cuadra, C. A., & Katter, R. V. (1967). *Experimental studies of relevance judgments: Final report*. Vol. I: Project summary, Vol. III: Complication of forms, guides, schedules, and instructions used in the studies. (NSF Report No. TM-3520/001/00). Santa Monica, CA: System Development Corp.
- Ellis, D. (1996). The dilemma of measurement in information retrieval research. *Journal of American Society for Information Science*.
- Fairthorne, R. A. (1969). Content analysis, specification, and control. *Annual Review of Information Science and Technology*, 4, 73-109.
- Fairthorne, R. A. (1972). Problems of data retrieval and dependent techniques. In A. I. Chernyi (Ed.), *Problems of information science*(pp. 98-110). Moscow: All-Union Institute for Scientific and Technical Information.
- Foskett, D. J. (1970). Classification and indexing in the social sciences. In *ASLIB Proceedings*, 22, 90-100.
- Foskett, D. J. (1972). A note on the concept of "relevance." *Information Storage and Retrieval*, 8(2), 77-78.

- Froehlich, T. J. (1971). --Towards an agenda for the 21st century. *Journal of the American Society for Information Science*, 45, 125-133.
- Gifford, C. & Baumanis G. J. (1969). On understanding the correlates of relevance judgments. *American Documentation*, 20, 19-26.
- Guertin, W. H., & Bailey, J. P., jr. (1970). *Information Science*. Ann Arbor, MI: Edwards Brothers.
- Hollnagel, E. H., & Mampower, J. (1980). *Human judgment and decision making*. New York: Praeger.
- Harris, R. J. (1975). *A primer of multivariate Statistics*. New York: McGraw-Hill.
- Hart, D. (1971). The structure of information science. *Journal of the American Society for Information Science*, 37, 615-625.
- Hert, C. A. (1991). *Understanding information retrieval interactions: Theoretical and practical implications*. London: Aslib.
- Ingersen, P. (1992). *Information interaction*. London: Taylor Graham.
- Ingersen, P. (1996). Cognitive perspectives of information retrieval interactions: A review. *Journal of Documentation*, 52(1), 3-50.
- Janes, J. W. (1991). Relevance judgments and the incremental presentation of information. *Information Processing & Management*, 27(6), 629-646.
- Kent, A., Taulbee, O. E., Belzer, J., & Goldstein, G. D. (Eds.) (1967). *Electronic information science*. Washington, D. C. : Thompson.
- Kuhlman, R. W. (1971). *Foundations of behavioral research (3rd ed)*. Fort Worth: Holt, Rinehart and Winston.

- Kim, J. O., & Mueller, C. W. (1978). *Introduction to factor analysis: What is it and how to do it?* Beverly Hills, CA: Sage Publications.
- Krzanowski, W. J. (1988). *Principles of multivariate analysis*. Oxford: Clarendon Press.
- Kuhlthau, C. C. (1993). *Seeking meaning: A process approach to library and information services*. Norwood, NJ: Ablex.
- Lonergan, B. J. F. (1970). *Insight: A study of human understanding*. New York: Philosophical Library.
- Marcus, R. S., Kugel, P., & Benenfeld, A. R. (1978). Catalog information and text as indicators or relevance. *Journal of the American Society for Information Science*, 29, 15-30.
- Maron, M. E. (1977). On indexing, retrieval and the meaning of about. *Journal of the American Society for Information Science*, 28(1), 38-43.
- Mizzaro, S. (1997). Relevance: The whole history. *Journal of the American Society for Information Science*, 48(9), 810-832.
- Park, T. K. (1992). *The nature of relevance in information retrieval: An empirical study*. Doctoral dissertation, School of Library and Information Science, Indiana University, Bloomington, IN.
- Park, T. K. (1994). Toward a theory of user-based relevance: A call for a new paradigm of inquiry. *Journal of the American Society for Information Science*, 45, 135-141.
- Patton, M. Q. (1990). *Qualitative evaluation and research methods* (2nd Ed.). London: Sage Publications.
- Rath, G. J., Resnick, A., & Savage, T. R. (1961). Comparisons of four types of lexical indicators of content. *American Documentation*, 12(2), 126-130.
- Rees, A. M., & Schulz, D. G. (1967). *A field experimental approach to the study of relevance assessments in relation to document searching*. Cleveland, OH: Case Western Reserve University.
- Rencher, A. C. (1998). *Multivariate statistical inference and applications*. New York: John Wiley & Sons.

- Renick, A., & Savage, T. R. (1964). The consistence of human judgments of relevance. *American Documentation*, 15(2), 93-95.
- Roberson, S. E. (1981). The methodology of information retrieval experiment. In K. Sparck Jones (Ed.), *Information Retrieval Experiment*. London: Butterworths.
- Robertson, S. E., & Hancock-Beaulieu, M. M. (1992). On the evaluation of IR systems. *Information Processing & Management*, 28(4), 457-466.
- Saracevic, T. (1969). Comparative effects of titles, abstracts and full-texts on relevance judgments. In *Proceedings of the American Society for Information Science*. Medford, NJ.: Learned Information.
- Saracevic, T. (1970). *On the concept of relevance in information science*. Doctoral dissertation, Case Western Reserve University.
- Saracevic, T. (1975). Relevance: A review of and a framework for the thinking on the notion in information science. *Journal of the American Society for Information Science*. 26, 321-343.
- Saracevic, T. (1996). Interactive models in information retrieval: A review and proposal. *Proceedings of the 59th Annual Meeting of the American Society for Information Science*, 33, 3-9.
- Saracevic, T. (1996). Relevance reconsidered '96. *Proceedings of the second International Conference on Conceptions of Library and Information Science*, 209-218.
- Saracevic, T. (1997). Users lost: Reflections on the past, future, and limits of information science. *Forum: A Publication of the ACM Special Interest Group on Information Retrieval*. 31(2), 16-27.
- Schamber, L. (1991). *User's criteria for evaluation in a multimedia information seeking and use situation*. Doctoral dissertation. Syracuse University.
- Schamber, L. (1994). Relevance and information behavior. *Annual Review of Information Science and Technology*, 29, 3-48.
- Schamber, L., Eisenberg, M. B., & Nilan, M. S. (1990). A re-examination of relevance: Toward a dynamic, situational definition. *Information Processing and Management*, 26(6), 755-776.

- Schamber, L. , & Bateman, J. (1996). User criteria in relevance evaluation: Toward development of a measurement scale. *Proceedings of the 59th Annual Meetings of the American Society for Information Science*, 33, 218-225.
- Smithson, S. (1994). Information retrieval evaluation in practice: A case study approach. *Information Processing & Management*, 30(2), 205-221.
- Singleton, R. A., Straits, B. C., & Straits, M. M. (1993). *Approaches to social research*. New York: Oxford University.
- Sparck Jones, K., & Willett, P. (Eds.). (1997). *Readings in information retrieval*. San Francisco: Morgan Kaufmann.
- Sperber, D., & Wilson, D. (1995). *Relevance: Communication and cognition* (2nd. ed). Oxford, UK: Blackwell.
- Swanson, D. R. (1986). Subjective versus objective relevance in bibliographic retrieval systems. *Library Quarterly*, 56, 389-398.
- Swanson, D. R. (1988). Historical note: Information retrieval and the future of an illusion. *Journal of the American Society for Information Science*. 39(2), 92-98.
- Tang, R., & Solomon, P. (1998). Toward an understanding of the dynamics of relevance judgment: An analysis of one person's search behavior. *Information Processing & Management*. 34(2/3), p. 237-256.
- Thompson, C. W. N. (1973). The functions of abstracts in the initial screening of technical documents by the user. *Journal of the American Society for Information Science*, 24 , 270-276.
- Tibbo, H. R. (1993). *Abstracting, information retrieval, and the humanities : Providing access to historical literature*. Chicago : American Library Association.
- Tenopir, C. (1985). Full-text database retrieval performance. *Online Review*, 9, 149-164.
- van Dijk T. A. (1979). Relevance assignment in discourse comprehension. *Discourse Processes*, 2, 113-126.
- Vickery, B. C. (1958a). Subject analysis for information retrieval. *Proceedings of the International Conference on Scientific Information*, 2, 855-865.
- Vickery, B. C. (1958b). The structure of information retrieval systems. *Proceedings of the International Conference on Scientific Information*, 2, 1275-1289.

A cognitive model of document selection of real users of information
 Doctoral dissertation, University of Maryland.

Wang, P. (1997). The design of document retrieval systems for academic users:
 Implication of studies on users' relevance criteria. *Journal of the American Society for Information Science*, 34, 162-173.

Webster's third new international dictionary of the English language unabridged. (1961).

strategies.
 In K. Borcharding, O. I. Larichev, & D. M. Messick (Eds.),
in decision making (pp. 159-171). Amsterdam: Elsevier Science Publishers.

White, M. D., & Wang, P. (1997). *Document selection and relevance assessments during research project*. College Park: University of Maryland.

Wildemuth, B. (1993). Post positivist research: Two examples of methodological pluralism. *Journal of the American Society for Information Science*, 44(4), 450-460.

Wilson, P. (1973). Situational Relevance. *Journal of the American Society for Information Science*, 24(4), 457-471.