DOCUMENT RESUME

ED 027 921 LI 001 344

By-Cuadra, Carlos A.

A Study of Relevance Judgments.

Pub Date [68]

Note-12p.; Final draft of talk for 1968 Annual Meeting of the American Psychological Association, San Francisco

EDRS Price MF-$0.25 HC-$0.70

Descriptors-*Evaluation Methods, Information Science, *Information Systems, *Relevance (Information Retrieval), *Systems Analysis

During the two-year Relevance Assessment Project, 15 studies were designed and carried out with over 500 subjects used as relevance judges. The subjects were librarians, information specialists, library science students and faculty, and graduate and upper division students in psychology. Materials for judging were selected for subjects according to their interests and backgrounds. Results of the many experiments conducted indicate that relevance judgments can be influenced by many factors including (1) the skills and attitudes of the judges used, (2) the documents and document sets used, (3) the particular information requirement statements, (4) the instructions and setting in which the judgments take place, (5) the concepts and definitions of relevance employed in the judgments, and (6) the type of rating scale or other instrument used to express the judgments. Findings indicate that relevance scores must be used with caution for system evaluation. A list of articles and reports on the project is included. (CC)

Relevance Assessment Project

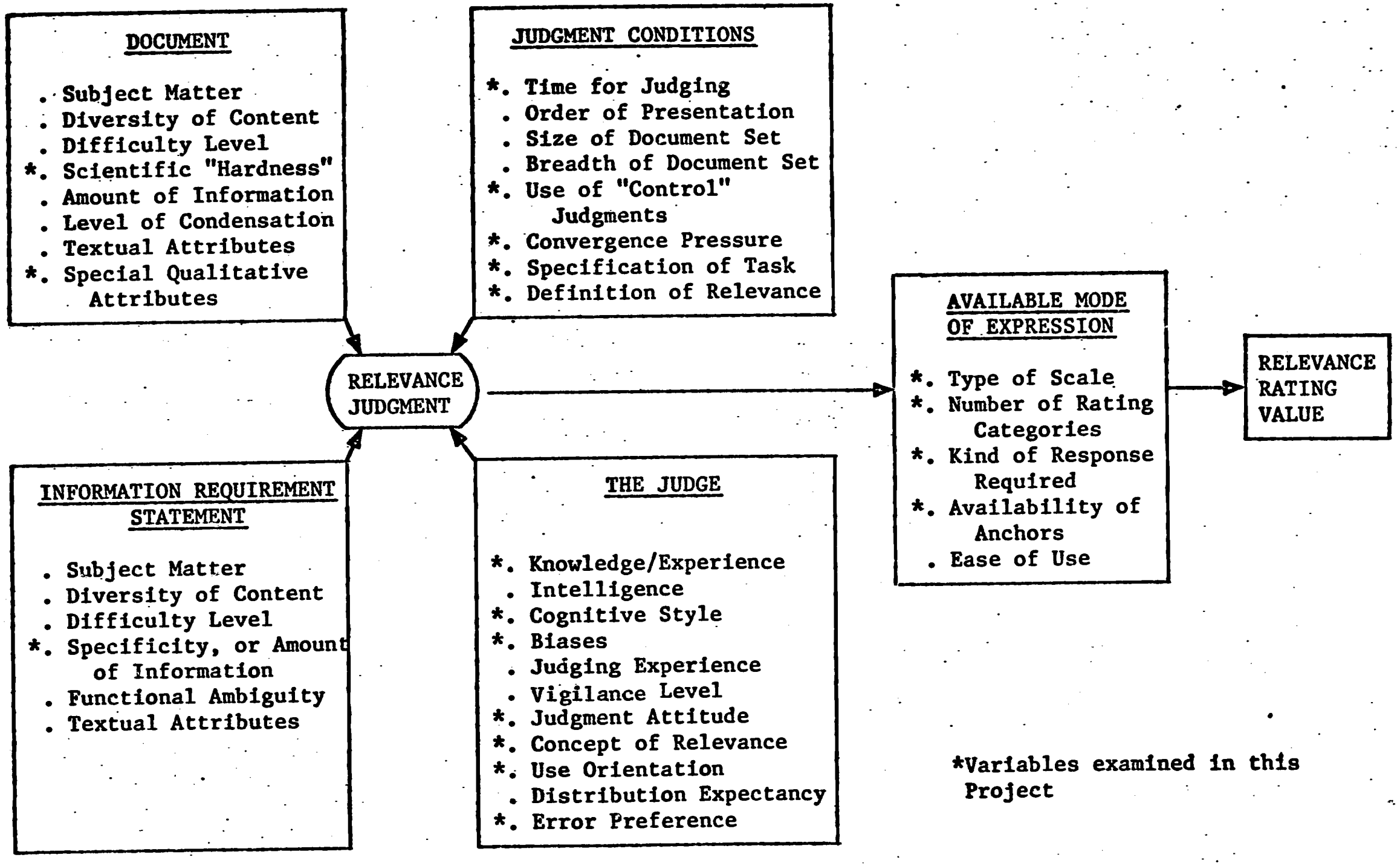
for

Symposium on Psychological Theory and Method in Information Science

1968 APA Annual Meeting

San Francisco, California

Dr. Carlos A. Cuadra
System Development Corporation
2500 Colorado Avenue
Santa Monica, California 90406



*Variables examined in this Project

Articles and Reports on the SDC Relevance Assessment Project

1. CUADRA, CARLOS A. Toward a scientific approach to relevance judgments. Presented at 33rd Conference of FID and International Congress on Documentation, Tokyo, 12-22 September 1967. (Preprint, 12 p.)

2. CUADRA, CARLOS A.; KATTER, ROBERT V. Opening the black box of relevance. Journal of Documentation, December 1967, vol. 23, no. 4, p. 291-303.

3. CUADRA, CARLOS A.; KATTER, ROBERT V. The implications of relevance research for library operations and training. Presented at Special Libraries Association Conference, Los Angeles, California, June 3, 1968.

4. CUADRA, CARLOS A.; KATTER, ROBERT V. The relevance of relevance assessment. In: Proceedings of the 30th Annual Meeting of the American Documentation Institute, New York, October 1967. Thompson, Washington, D. C., 1967, p. 95-99.

5. KATTER, ROBERT V. The influence of scale form on relevance judgments. Information Storage and Retrieval. Pergamon Press 1968, vol. 4, p. 1-11.

6. SYSTEM DEVELOPMENT CORPORATION. Experimental studies of relevance judgments: final report. By Carlos A. Cuadra, Robert V. Katter, Emory H. Holmes, and Everett M. Wallace. Santa Monica, Calif., 30 June 1967, 3 vols. (Report nos. TM-3520/001/00; TM-3520/002/00; TM-3520/003/00).

A STUDY OF RELEVANCE JUDGMENTS*

Before discussing our research, I'd like to comment that I'm very glad to be talking before an APA audience again, after an interval of 13 years. The last time was in 1955, when I reported on a study that Dr. William Albaugh and I had conducted on the communicability of clinical psychology reports. I began thinking about that study again a few weeks ago as I tried to remember what APA conventions and APA audiences were like.

When I left the field of psychology--at least as an active worker--some twelve years ago, I didn't realize that I had any interest then in what is now being called information science. Yet as I look back, I realize that even then I was interested in the flow of information and in professional communication. The study on psychological reports, which I associated only with clinical psychology work in my distant past, suddenly became highly relevant to my current work in the field of information science and technology.

Thinking about that study also reminded me that there was a time when we psychologists were noticeably thin skinned about the profession. One of the rather startling findings of our early study was that about half of the messages that clinical psychology report writers were trying to convey to the report readers either did not get through or got through in a highly distorted form. The reason I undertook the study was because of a gnawing feeling that the psychiatrists and other readers of the reports that I and other staff members and trainees were so

*This is final draft of talk for 1968 Annual Meeting of the American Psychological Association, San Francisco, California.

painstakingly creating might not really understand what we were saying. We set about checking this hunch by developing a series of questionnaires for each of several psychological diagnostic reports, then asking psychiatrists, other physicians, nurses, staff psychologists, and psychology trainees to study the reports and indicate, by means of the questionnaire, what the writer had told them.

Each questionnaire item was multiple choice and contained, in addition to one item that was almost directly lifted from the report or was a fairly direct paraphrase, one or more wild, totally wrong interpretations of what the author thought he had said. (The author incidentally, provided the criterion judgments for our test.) The result was about 50% "correct" responses. The psychology staff did a little better than psychology trainees, who in turn did a little better than student nurses and psychiatrists.

This paper reporting on these results was accepted for the APA convention in San Francisco, and the APA publicity people planned to have press releases and interviews. These were subsequently called off, because of the feeling that the results of the study might be misinterpreted and misused by those hostile to psychology. I have occasionally wondered if half a generation of clinical psychologists since that time has continued to turn out acres of reports that half of a generation of psychiatrists still don't understand. (I hope, incidentally, that APA is more comfortable and less thin skinned about its image now than it was 12 years ago.)

All of this brings me circuitously to the subject of relevance.

I should begin by trying to indicate why the study of relevance judgments has seemed important to the field of information science.

The large-scale use of data processing equipment and procedures in libraries and document-handling systems began in the 1950s. Within a few years, the appreciation of a number of technical problems in such use, together with the fairly sizable costs of automated or semiautomated information retrieval systems, helped to awaken an active interest in system effectiveness and in means of measuring it.

Almost from the beginning of serious work on the evaluation of information retrieval systems, which began in 1953 at the Library of Congress, attempts to provide adequate criteria of evaluation were dependent on and bedeviled by the concept of "relevance", relevance usually referring to a relationship between some kind of information need and some kind of system output, such as a document. The evaluation study in the Library of Congress proved inconclusive, incidentally, at least in part because the evaluators and the Library of Congress personnel could not agree on which documents were <u>really</u> "relevant."

Evaluation studies generally followed the same general pattern: the holdings of a system were searched, in response to some kind of inquiry, and a subsequent judgment was made as to whether the resulting outputs were "relevant" to the inquiry. On the basis of these judgments, various scores were computed to express the system's retrieval performance. After some years, these scores began to be discussed by some workers with a great deal of reverence. One such score was named "recall;" it refers to a ratio of two numbers: the number of "relevant" documents produced by a retrieval system over the total number of "relevant" documents actually

in the system's store. It was quite common to have people write or say such things as "Information System X is performing at 80 percent recall, while Information System B is performing at only 75 percent recall."

Most of the workers engaged in retrieval system evaluation from 1953 to the present have had relatively little interest in relevance judgments per se. They have been interested in them primarily as a criterion by which to evaluate manual or computer-based searches, or comparisons between them. These workers, of course, have been aware of disagreement among judges, but they have tended to consider such disagreement largely as an irritant, to be stamped out or bypassed as quickly as possible, rather than as a phenomenon worthy of interest in its own right. Thus, in spite of the reliance in system evaluation on the notion of a "relevant set of documents," the relevance process itself has largely been treated as a "black box," and there has been very little effort to understand either what goes on inside the box or how variations in the judgments might lead to variations in the identification of the relevant set of documents. This is somewhat analogous to a situation in psychology where we used a test for the selection of personnel without knowing what the test measures, how the test items are interrelated, what factors cause variations in test scores, or what relationships individual test items and test dimensions have to specific aspects of job performance.

Against this background, and against a backdrop of frustration and disagreement about the validity and implications of evaluation studies involving relevance, we began a project under support from the National Science Foundation, to develop some empirical information about human judgments of relevance. Our study at System Development Corporation began at the same time as a companion project at Case Western Reserve University, which Dr. Schultz will describe a little later.

In planning our project, we made two important methodological choices. First, we focused on relevance as a relationship between a document and some public, visible expression of an information need, to wit, a written statement describing an information need. Second, we chose to follow an experimental, laboratory approach,to permit greater control over the important variables. It was our feeling at the outset that we might eventually be dealing with dozens or scores of variables, each of which could be measured in many ways. It seemed important, therefore, to operate in circumstances that would help us to see as clearly as possible the effect of particular variables and the usefulness of particular measures.

Prior to any experimentation, the Project staff developed a list of variables that might be contributors to variations in relevance judgments. Since there was little empirical evidence related to any of these, the list was based almost entirely on a priori considerations. Groups of variables relating to five aspects of relevance judgments were identified: (1) Documents, (2) Information Requirement Statements, (3) the Judge, (4) the Judgment Conditions, and (5) the Available Mode of Expression. Within these groups a total of 38 variables was listed, as shown on the second page of the handout.

During the two years of the project, we examined almost half of the 38 variables on our list. Fifteen studies were designed and carried out, using over 500 subjects as relevance judges. The subjects were librarians and information specialists, library science students and faculty, and graduate and upper division students in psychology. Materials for judging were selected and/or created in accordance with particular experimental objectives and the backgrounds of the judges.

The attempt to look at almost 20 variables was premised on the fact that many of these variables were likely to be related to each other. Any single experiment--in fact, any relevance judging situation of any kind--must use a particular set of documents or representations, a particular set of information requirement statements, particular judges, particular judgment conditions, and particular modes of expressing the relevance judgments. Yet each of these is itself a potential source of variability. Therefore, one cannot generalize the results of any single experiment, no matter how well it might be done, because the influence of other variables may not be known. For this reason, it was methodologically preferable to attempt a first-round assessment of many variables, rather than an intensive study of a single variable or small group of variables, in isolation.

There are several detailed reports on our studies, noted on the last page of the handout, and I won't attempt to summarize them here. I would like to mention the findings from one of the studies having to do with the negotiation process between an information user and a librarian or information specialist.

This study looked at what we call the "implicit use orientation" of the user. By use orientation, we mean the user's expectation regarding the way in which he will use the information. For example, he may be trying to compile an exhaustive bibliography; or to identify articles that contain specific bits of information of some immediate practical use; or to get articles that serve no particular practical use but may have idea-stimulating value. You have been exposed to and experienced such orientations; what we wanted to do is see whether they influence judgments of document relevance.

To do this, we exposed about 150 judges to a set of 9 documents and 8 information requirements statements. Then, after eliciting relevance judgments, we got a second set of judgments under different conditions. For the second set of conditions, we, in effect, told the judges more about the users and their use orientations, and we asked the judges to consider themselves to be acting as agents for the users as they made their judgments on the documents and information requirements statements. For this part of the experiment, we divided our judges into 14 groups, each of which was given a different use orientation. We then compared the resulting data with the judgments that the same judges had made earlier, without any special use orientation. The results showed that each of the 14 use orientations we imposed altered the relevance scores that the judges assigned to documents. A document what would be accorded high relevance for a bibliography orientation might be given low relevance for some other kind of orientation.

What the study showed, among other things, is that relevance scores are very slippery. Documents clearly have no inherent, unchanging relevance to information requirements statements; the relevance values attributed to them really depends, in part, on how the documents are going to be used.

This was only one of many experiments undertaken during our two years of work. The results from all the studies, taken together, show that relevance judgments can be influenced by many factors: the skills and attitudes of the particular judges used, the documents and document sets used, the particular information requirement statements, the instructions and setting in which the judgments take place, the concepts and definitions of relevance employed in the judgments, and the type of rating scale or other instrument used to express the judgments.

I don't think these kinds of findings would surprise many people trained in psychology. We have all been trained to think of behavior in terms of many variables, some of them highly complex and obscure. Yet the kind of research I have described is relatively new in the information science field and even though psychologists would fully expect experimental results to hinge on the kinds of judges, documents and other variables involved in the experiment, neither they nor anyone else has been in a position to say how these variables behave in the document-judging situation.

From the standpoint of system evaluation--which was our orientation at the outset of the project--our findings cast serious doubt on the unquestioning use of relevance scores as table criteria for system or subsystem evaluation, because these scores are likely to be artifacts of particular systems and of the particular conditions of relevance measurement. Thus, they may not deserve the aura of quantification, validity and stability that they currently enjoy. Our findings also suggest that the use of single figures of merit (for example, "our system has 80 percent recall") can be quite misleading in comparisons between different information systems or, indeed, under any circumstances where the sources of variability mentioned above have not been taken into account and controlled. Too, even if one were to develop stable and meaningful figures of merit for information systems, then what does he do to improve the system? It is obvious that system improvement rests not on overall figures of merit but on sensitive diagnostic information on particular aspects of system performance, such as Dr. Katter discussed in his paper. The importance of relevance work is not that it will provide better figures of merit, but that it will help us to understand better the interface between the information system, on the one hand, and the user or intermediary, on the other. Such understanding is an absolute requirement for effective diagnosis.

Our studies have provided results that I know are frustrating to some workers in the information science field, and there has been some feeling expressed that relevance judgments are not only suspect but unworthy of study; therefore we should dispense with them. That, to me, is a very cheap way out of a predicament. You will recall that many years ago, in the field of psychology, there was a widespread revolt against subjective phenomena--in part because of the same frustrations some of us have experienced with relevance. The outcome of the revolt was that, for a time, psychologists devoted their attention not to what was important, but to what was measurable. This is a surefire approach to a certain kind of respectability, which information scientists now desire as much as psychologists did then, but it risks losing the baby with the bathwater. I believe that, when information scientists fully accept the fact that relevance phenomena are complex and slippery, they should not take the easy way out of simply turning their back on such phenomena. Relevance judgments, however disguised and however renamed, are indispensable aspects of our field, and part of the challenge is to admit their complexity, to start trying to learn what they are about, and to begin building better, and less elastic, rulers to measure them.