Inform. Stor. Retr. Vol. 4, pp. 1–11. Pergamon Press 1968. Printed in Great Britain

THE INFLUENCE OF SCALE FORM ON RELEVANCE JUDGMENTS

R. V. KATTER*

System Development Corporation, Santa Monica

Summary—This paper reports the results of two studies. The first compared ranking and category rating procedures for measuring relevance of documents to information requirement statements. The comparison measure was number of reversals; the condition where document-requirement pair A is measured as *more* relevant than pair B by one procedure, and as *less* relevant than pair B by the other. As compared to category rating, ranking produced three times the expected number of reversals. The results are explained in terms of a "cascaded distortion process" that can affect any procedure which arbitrarily restricts distribution shape.

The second study compared the stimulus range and anchoring sensitivities of a nine-point category scale and a magnitude-ratio scale procedure. Results from the category scale were more consistent and more as predicted. Magnitude ratio results were distorted by unrepresentative scale moduli selected by about one-sixth of the judges, a condition which may be correctable. Suggestions for improved anchoring procedures are discussed in light of the findings for the anchoring treatment.

I. THE RELEVANCE PROBLEM

RECENT widespread activity in the area of automated information systems has brought about a concomitant interest in problems of system evaluation. This interest reflects not only an understandable concern about the value of automated retrieval systems, but also a growing maturity and scepticism in the information processing field with regard to claims of system capabilities and performance.

Nearly all studies that attempt to evaluate the effectiveness of information retrieval systems or the effectiveness of their indexing subsystems—generally conceded to be the most critical element—have relied heavily on the notion of a "relevant set of documents". The identification of such relevant sets of documents has typically involved panels of technically competent judges rating a number of documents against a number of information-requirement statements, with the resulting ratings serving as the basis for evaluation of the output from searches in a retrieval system.

A recurring finding from studies involving relevance judgments is that the inter- and intra-judge reliability of relevance judgments is not very high. Most of the researchers who have been confronted with this fact of unreliability have treated it largely as an irritant to be stamped out as quickly as possible, rather than as a phenomenon worthy of interest or study. The primary interest in relevance in system evaluation studies has been in obtaining criterion measures, rather than on understanding the judgmental phenomena. In addition to the lack of interest in judgment behavior itself, many of the studies involving relevance judgments and scores appear to have methodological shortcomings, in some cases so severe that findings that purport to provide quantitatively sound measures of system per-

* The author wishes to acknowledge the contributions made to the writing of this paper by Carlos Cuadra, Emory Holmes, Everett Wallace, and Robert McCornack, all of the System Development Corporation, Santa Monica.

formance could justifiably be suspected of reflecting primarily the differences in the conditions of measurement. It has been difficult to tell, however, to what extent this suspicion is justified, because of the absence of *empirical* data on sources of variability in relevance judgments.

In 1965, the National Science Foundation initiated support for two projects, for the purposes of developing empirical evidence on the sources of variability in relevance judgments and for determining how such variability can be dealt with. A previous report [1] dealt with the rationale for intensive study of relevance judgments and described results from a study of judge's attitudes toward the judging task. That study demonstrated that the relevance score assigned by a judge is, in part, a function of his conception of the use to which he imagines he will later put the document. It was also demonstrated that the relevance scores assigned by judges to document-requirement pairs can be manipulated (raised or lowered) without difficulty through particular instructions to the judges. Details of these and several other experiments are provided in progress reports from the project [2] [3].

II. PURPOSES OF PRESENT STUDY

The studies described here were intended to illuminate one major class of variables influencing relevance scores: the form of the response elicited from the relevance judge. Evaluation studies employing relevance measures of various kinds have used a variety of scale forms. Such studies have not justified their selections of scale form on empirical grounds; indeed, it would be difficult to do so, because empirical evidence on the influence of scale form is virtually nonexistent in the field of documentation. A few recent examples from among many studies using contrasting scale forms for various variables are found in RIDDLE, HORWITZ and DIETZ (1966) [4], who used rankings as final comparison-measures for investigating the output of a machine-based method for relevance feedback to improve querying techniques; a report by GIULIANO and JONES [5] uses retrieval rank as the system performance measure and a 5-category rating scale as the evaluation measure; STILES [6] used retrieval rank as the main output measure; and CLEVERDON, MILLS and KEEN [7] employ a 5-category rating scale of relevance as a criterion measure.

Extensive work on scaling in the field of measurement psychology suggests that the form of the scale used can have a sizeable impact on the values obtained, and also that some scales may be more appropriate than others for measuring certain variables. The two studies reported here were aimed at an initial exploration of these possibilities as applied to relevance judgments. The first study, which compares ranking and rating, attempted to clarify the potential impact of procedures, such as ranking, that restrict the shape of the distribution of relevance scores. The second study attempted to test a possible procedural improvement over the category scale commonly used for relevance judgments.

III. RANKING VERSUS CATEGORY SCALES

At times, the words "rating" and "ranking" have been used in the system evaluation literature in such a way as to appear interchangeable, which—of course—they are not. A ranking procedure forces (i.e. necessarily produces) a rectangular distribution of scores which may have a variety of distortion effects, as compared to rating scales. The magnitude of the distortion of judgments depends upon how different the shape of the distribution would be if allowed to vary freely. The nature and importance of such distortions are often not recognized. Two kinds of distortions can result from such restrictions: first, an apparent judgmental distinction will sometimes be scored between two objects (e.g. between two

document-requirement pairs) when in fact the distinction is not actually supportable in the judge's experience. Second, a distinction between two objects will not be scored in instances where the judge perceives a distinction. Distortions of these kinds may have no important effect on information system operations, so long as they do not alter the *ordering* of judged relevancies. However, to the extent that orderings are *reversed*, there will be effects on information retrieval system operations.

A *reversal* can be seen in the case where, in one series of obtained judgments, document-requirement pair A is depicted as more relevant than pair B, and in the other series of obtained judgments, pair B is depicted as more relevant than pair A. Because of its direct implication for information system operations, obtaining percentage of reversals is an especially illuminating way to depict the amount of disagreement or distortion occurring between distributions of judgments obtained in different ways on the same set of pairs. However, the obtained percentage of reversals should actually be expressed as a multiple of the percentage of reversals obtained as a function of the unreliability of the judgment process itself (i.e. as a multiple of the percentage of reversals obtained in an immediate test-retest reliability study, using the same judges, materials, and instructions for both tests). For a typical test-retest reliability of 0·75, the percentage of reversals had a value of about 5 per cent in one sample of our data on which reversal counts were made.

There are two ways to force the shape of a distribution. One is to prescribe the relative frequencies of judgments that must fall into each category, so that the distribution is normal, rectangular, or some other predetermined shape. The other is to rank the objects rather than rate them, thus forcing a rectangular distribution. To assess the distortion that might occur in ranking, a comparison was made between results obtained from rating and ranking versions of the same judgments.

A. *Study methods employed*

In the first study reported, 12 advanced psychology students first *rated* the relevance of each of eight short information-requirement statements to each of nine abstracts of journal articles in the area of social psychology and personality theory. The relevance judgments for the 72 (8×9) document-requirement pairs were expressed on a 9-point category scale of overall relevance, shown in Fig. 1. After a 15-min break, the judges *rank-judged* the same 72 pairs, using the instructions shown in Fig. 2.

In the completion blanks on the following pages, insert the number of the statement below that best expresses your judgment of the relevance of the article to the particular requirement statement.

1—"has *absolutely no* relevance".
2—"has *near minimum* relevance".
3—"has *weak* relevance".
4—"has *noticeable* relevance".
5—"has *very noticeable* relevance".
6—"has *strong* relevance".
7—"has *very strong* relevance".
8—"has *near maximum* relevance".
9—"has *maximum* relevance".

FIG. 1. Category scale of overall relevance.

You are to imagine that you are a person dealing with an information retrieval system of a kind that must always be able to "break the tie votes" in its processing operations. Therefore, the system cannot tolerate the same relevance value to be assigned to any two abstract-requirement pairs that happen to share either an abstract or a requirement in common. Thus all your judgments must be made in a different form, that is, a *ranking* form. You will try to judge the relevance of articles accordingly.

JUDGMENT PROCEDURE

For each abstract:

1. Place a (1) in the cell of the statement for which it has *most* relevance.
2. Place an (8) in the cell of the statement for which it has *least* relevance.
3. Place a (2) in the cell of the statement for which it has *second most* relevance, etc.
4. Continue in this manner until all cells for an abstract are ranked.
5. Then go on to the next abstract.

FIG. 2. Instructions for ranking procedure.

To compare the results from ranking with those from rating, it was necessary to compare each "object" (document-requirement pair) with every other object to determine whether one of the two had a larger relevance value assigned to it. With N objects there are $(N^2 - N)/2$ such comparisons for the 12 judges. To make the necessary comparisons, a special computer program was devised, which scored reversals.

B. *Results*

Of the 30,672 possible reversals, there were in fact 5066, or 16·5 per cent. On the basis of the test-retest figure of 5 per cent mentioned earlier, this is about three times the number that should be expected. The increase in reversals can be attributed to differences in the scale forms. About half of the reversals had a value of 5 or greater, where the minimum possible value was 2 and the maximum was 15. (The minimum reversal value of two was obtained when two pairs were one scale unit apart in one direction by one method, and one scale unit apart in the opposite direction by the other method.)

These results are better understood from graphic displays. In Fig. 3, the actual distributions of values obtained for the ranking and rating methods are shown. It is clear that for the rating distribution (which had no restriction on its shape), relevance values

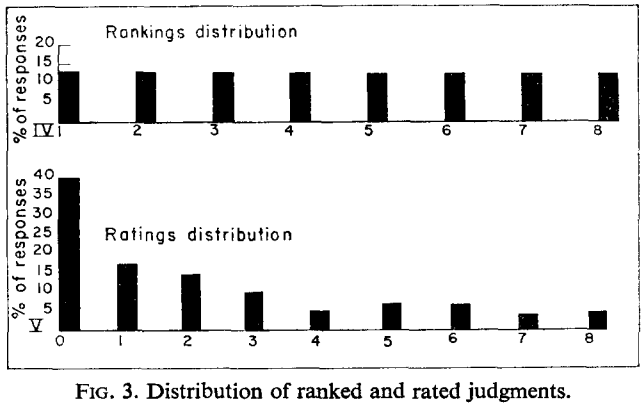

FIG. 3. Distribution of ranked and rated judgments.

tended to pile up near the bottom or low-relevance end of the scale. The shape restrictions on the ranking distribution pulled or distorted these values into a rectangular shape.

The mechanism of this distortion is quite simple and might be called a "cascaded distortion process"; in the *rating* procedure, any value can be used as many times as the judge finds appropriate. In the *ranking* procedure, however, once a value has been used it cannot be used again, and the judge is forced to assign the nearest available value. As slots get used up, the judge must assign values farther and farther away from the ones he would use if his choices were unrestricted. The effect of this process depends on the shape of the unrestricted distribution, and on the particular sequence of judgments in the ranking procedure. The cascaded distortion effect for this study can be seen in Fig. 4.

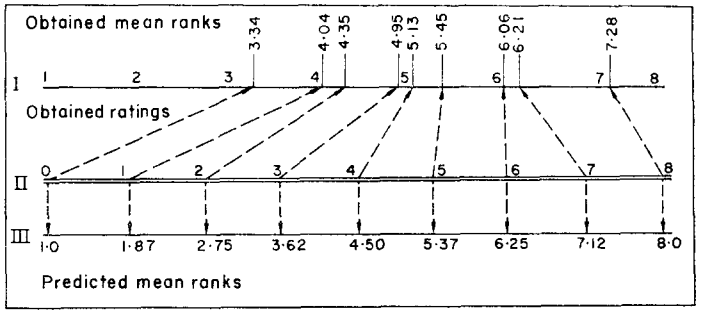

FIG. 4. Distribution between rated and ranked judgments.

In Fig. 4, row III shows the mean *rank* values that would be obtained for objects *rated* at the corresponding values in row II if there were a perfectly reliable judgment process and no systematic distortion introduced by the distribution restrictions of the ranking method. Row I shows the *actual* mean ranks obtained for objects rated at corresponding values in row II. Notice that the cascaded distortion process has compressed the mean rankings from both ends, with the effect being more severe at the bottom, due to the larger number of cases falling there.

These results show that restrictions on the shape of the distribution can do violence to the underlying judgmental phenomena, to the extent that the restrictions do not match the judgmental facts. The resulting distortions can seriously mar the comparability of experiments, as well as change the outputs of retrieval systems which use rank-judged data, or weights based on such judgments, for comparing documents as answers to information requirements.

IV. CATEGORY VERSUS MAGNITUDE-RATIO SCALES

The study reported above indicates that, in comparison with category rating scales, ranking procedures may introduce significant distortions into relevance judgments. In turn, there is some evidence recently summarized by STEVENS [8] that under certain circumstances category scales introduce more distortion than magnitude-ratio scales.

In the magnitude-ratio scale procedure, the judge selects any arbitrary number to express the magnitude of his judgment of the first object (i.e. document-requirement statement pair), and then expresses all subsequent judgments as *ratios* of the first magnitude: "half as great", "twice as great", etc. Whereas a category scale with values running, for example, from one through nine allows the expression of a maximum magnitude nine times as great as the minimum expressable magnitude, the magnitude-ratio procedure

involves no inherent restriction on either the ratio between the smallest and largest numerical values the judge may use, or on the number of different categories he may use.

Stevens cites several studies in which different judgmental continua that appeared to have about equal ranges when subjected to category scales displayed widely disparate ranges when magnitude-ratio scales were used. These results are presumably explained by the fact that the category scaled restricts the expression of judgments to the range and types of the prechosen values, even though the actual judgmental magnitudes may have a much larger range.

Stevens also suggests that, as compared to category scale ratings, ratings made by the magnitude-ratio estimation procedure may be less affected by biases in the particular sampling of objects to be judged. For example, suppose one group of judges is given a sample of objects that have been preselected for high values on the attribute to be judged, while another group is given objects preselected for low values on the attribute. If both groups of judges use a 9-point category scale to express their judgments, the lowest-valued objects in *both* the high- and low-valued samples might receive a scale value rating of "1", while the highest-valued objects in both samples might receive a value of "9". Stevens' position is that the category scale's dependency on context would *tend* more to operate in this way to obscure the differences between the samples of objects, and reflect mainly the perceived differences between objects within each separate sample.

If, instead, both groups of judges had used a magnitude-ratio estimation procedure, Stevens' claim is that there would be more of a tendency for the rating values to preserve the preselected differences in the two samples of objects. In that case, judges would tend to assign *lower* numerical values to the lowest-value objects of the *low*-value sample than to the *lowest*-value objects of the *high*-value sample. They would also tend to assign *higher* values to the highest-value objects of the *high*-value sample than to the highest-value objects of the *low*-value sample. The claim for less context dependency of the magnitude-ratio procedure can be subjected to empirical test in the manner indicated in this example.

Another procedure that can be used to reduce the effects of biases in the sample of objects is by the use of an "anchoring" example presented to the judge. He is told where the anchor object fits in the entire range of values of the attribute being judged, and tries to use this information to modify his rating values so that they reflect more accurately the positions of the judged objects within the entire range of possible values. Presumably, a scale procedure supposed to be more sensitive to the actual ranges of judgmental magnitudes, such as the magnitude-ratio procedure, would also be more sensitive to the effects of introducing "anchors" for judgment. Therefore the aim of the present study was, first, to see whether a magnitude-ratio scale would show more range-sensitivity for relevance judgments than a 9-point category scale, and second, to test the idea that the magnitude-ratio scale might also be more sensitive to the effects of an anchoring stimulus.

A. *Procedure*

To serve as judgmental stimuli, two sets of 12 abstract-requirement statement pairs from the field of psychology were selected from among a much larger set of pairs, on the basis of the mean relevance ratings assigned to the pairs of the larger set in an earlier study. A wide-range and a narrow-range set were selected in the manner indicated by Fig. 5.

In Fig. 5, the horizontal line represents a scale of previously rated relevance, from low to high. The narrow-range set consisted of 12 pairs that had all previously been rated as highly relevant. The wide-range set of 12 pairs was taken half from each extreme of the

pre-rated continuum. In addition, a single anchoring pair was taken from the middle of the continuum.

The two sets were each rated by both the category scale and the magnitude-ratio scale, both with and without the anchor pair. This provided the eight experimental conditions shown in Fig. 6. The judges for each condition consisted of separate samples of between 10 and 12 advanced psychology students. The 9-point *category* scale procedure was the same as that already described earlier in this paper. For the *magnitude-ratio* procedure, judges received the instructions shown in Fig. 7.

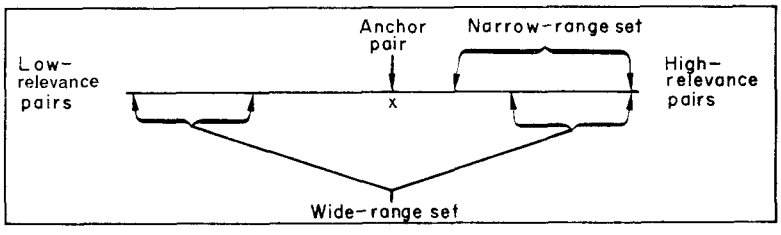

FIG. 5. Selection basis for judgmental stimuli.

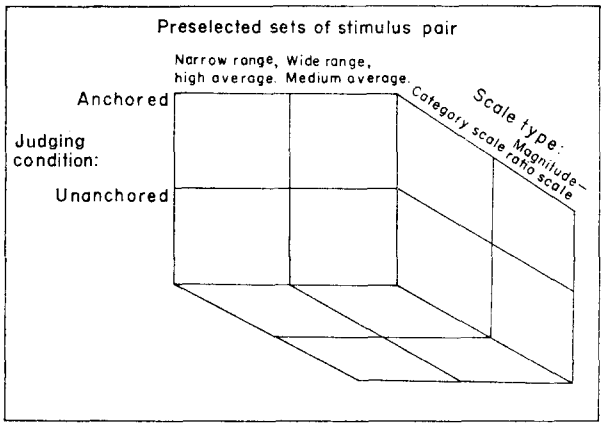

FIG. 6. Experimental design for scale sensitivity study.

The procedure is based on the observation that different individuals feel most comfortable using different-sized numbers to express their quantitative judgments. Follow these steps in assigning relevance values:

1. Select an abstract-requirement pair that seems to be intermediate in relevance *as compared to the other pairs in your booklet.*
2. Assign this selected pair any number you like to express its relevance value (try to select a number between 10 and 100).
3. Assign all other values *with respect to* this first number you have chosen. That is, if the next pair to be judged seems *twice* as relevant, assign it *twice* as big a number. If it seems *half* as relevant, assign it *half* as big a number. If it seems *four* times as relevant, assign it *four* times as big a number, etc.
4. Follow the above procedure to assign numbers to all ten judgment pairs.

FIG. 7. Instructions for magnitude-ratio procedure.

Since this type of judgment task is probably quite new to you, it is valuable to help you quickly gain some perspective on the *overall* range of possible relevance values between articles and requirement statements. Below is an example which you may use as an "anchor" in making your judgments. This particular article-requirement pair has been consistently judged as having a *medium* degree of relevance as judged among the *general run* of article-requirement pairs. (Your particular set of pairs may or may not fully cover the range of the general run of such pairs.) Study the example below carefully and refer back to it as often as you like in making your judgments.

ARTICLE:

A Syllabus of the Exoskeletal Defenses

This theoretical paper generalizes certain diverse forms of behavior as exoskeletal defences—a modification or enhancement of the body that subsidizes the psychological integrity of the person. The maneuver is one of armoring the self by strengthening the body (cosmetics, clothing, tattooing), or equipping the body with indices of invulnerability (amulets, tattoos, possessions, etc.). Exoskeletal defenses are interpreted as emerging from a perception of externality as threatening, as being consciously adopted to meet this press, and as having the advantage of being socially acceptable. They are contrasted with the classical defenses.

REQUIREMENT STATEMENT:

Need information on the social values aspects of personal empathy processes that may result in different sensitivities to various kinds of social influence.

FIG. 8. Instructions for anchoring treatment.

For the judgment conditions made *with anchor pair*, the instructions shown in Fig. 8 were added to the instructions for both the category scale procedure and the magnitude-ratio procedure.

B. *Analysis of data*

From the relevance ratings of each judge, three measures were calculated: the *mean* of all his relevance ratings, the *standard deviation* of his relevance ratings, and the *mean correlation* between his profile of ratings and the profiles of each of the other judges in his treatment group (between-judges agreement score). For each of the two scale types, a two-by-two analysis of variance was done for each of the three measures just described. For these six analyses, one treatment variable was the narrow- and wide-range stimulus sets, and the other was the anchored and unanchored conditions.

1. *Results of analysis of variance for means data.* Since the narrow-range stimulus set consisted of all high-relevance pairs, and the wide-range set consisted of half high and half low relevance pairs, it was predicted that ratings would average higher for the narrow set than for the wide set. This held true for the *category* scale results, beyond the 0·05 level of significance. For the magnitude-ratio scale results, however, the trends were in opposite directions for the anchored and unanchored treatments, and were not significant.

Since the function of the anchoring item was presumed to be that of making judgments more absolute, i.e. less dependent on the sampling bias of particular items presented, it was expected that the differences in predicted means described in the previous paragraph (narrow set with higher mean) would be accentuated for the anchored treatments as

compared to the unanchored treatments. There was a slight trend in the expected direction (nonsignificant) for the *category* scale results, but no trend for the *magnitude-ratio* scale. Thus the anchoring item did not work as expected. However, the anchor did have one clear effect for the category scale results: the mean rating was significantly higher for the anchored treatment than for the unanchored. The higher ratings assigned to the anchored set were nearer to the values they had previously been assigned in the preselection study, when these pairs had been presented in the context of a much wider-ranged set of pairs. In this sense, the anchor items showed interpretable anchor effects, but for the *category* scale only.

2. *Results of analysis of variance for standard deviations data.* It was also expected that the standard deviations for distributions of ratings made from the wide-range stimulus set would be greater than for the narrow-range set. This was true (beyond the 0·05 level of confidence) for the *category* scale, but the relationship was insignificant and reversed from the prediction for the *magnitude-ratio* scale.

3. *Results of analysis of variance for between-judges agreement data.* Predictions for the between-judge agreement values involved the following reasoning: As compared to the narrow-range set, pairs of pairs from the broad-range set should have had many more large, easily discriminable differences in degree of relevance between one pair and another. As a result, there should be more instances of agreement between judges as to the *ordering* of pairs in these large-difference pairs of pairs. The increased percentage of agreement, then, should first produce higher average between-judge intercorrelations of ratings for the wide-range set of pairs. Second, the expected difference should be greater for the magnitude-ratio scale than for the category scale, because of the presumed greater context-independence provided by the magnitude-ratio procedure.

The first hypothesis, that between-judges agreement would be greater for wide-range than narrow-range treatment groups, was confirmed beyond the 0·01 level for the *category* scale, and beyond the 0·001 level for the *magnitude-ratio* scale. The greater significance for the latter scale is in line with the second hypothesis of greater context-independence for the *magnitude-ratio* procedure.

The fact that the interjudge agreement results were the only ones that confirmed predictions for the magnitude-ratio scale is very enlightening. They indicate that when each separate judge's rating values were *standardized* (by being processed by a product-moment correlation computation, thereby being rescaled to effectively produce equal means and sigmas for all judges), the magnitude-ratio results appear as orderly as the category scale results. These findings provide a clue to what went wrong with the magnitude-ratio scale; it was the wide disparity between judges on their individual means and sigmas for magnitude-ratio ratings that produced the apparent disorder.

Acting on this cue, the raw rating values for magnitude-ratio results were examined. It was found that the *disparate* judgment values of only 6 of the 42 judges accounted for most of the unexpected results for that scale. (As one example of disparate results, in one treatment involving 11 individual judges, 9 of the means ranged between 28·7 and 67·2, while one judge's mean had a value of 8·8 and another had a value of 218·1.) Thus, in Stevens' terms, about one out of six judges selected a highly unrepresentative "modulus" (range of numerical values) for expressing his magnitude-ratio judgments. In a more recent article, STEVENS [9] discusses the problems created when judges choose unrepresentative moduli for expressing the judgments and suggests several possible methods of pretraining that might help to avoid this problem.

V. DISCUSSION AND FINDINGS

Together, the results from the two studies suggest that, for the present, category rating scales must be considered superior to both ranking and magnitude-ratio procedures for the expression of relevance judgments. This conclusion in no way fixes the decision for the possible future, however. The magnitude-ratio scale was outperformed by the category scale in part because of its apparent inherent limitation—excess sensitivity to each particular judge's idea about the number system—and in part because the experimental method used was not, after the fact, optimum. The case for the modulus selection unreliability of the magnitude-ratio scale is not settled for relevance judgments. Stevens suggests that ways may be found for standardizing the modulus selection of different judges. If so, the superior range-sensitivity of magnitude-ratio type procedures might yet be established for relevance judgments.

Regarding the experimental procedures, it is evident first that the single-item anchor at the midpoint was not adequate to provide the desired freeing of the judgments from the context effects produced by biased sampling of document-requirement pairs. Adding additional anchors at the extremes seems a possible solution. Second, there appears to be evidence that the range of rating values assigned to a set of such stimuli is partly a function of the number of discriminably different values of stimuli presented, i.e. the "spaces" in the range of stimuli apparently need to be "filled in" by just discriminable steps. This observation harks back to Thurstone methods for scale construction, and suggests that for sufficiently important areas of application, effective sets of anchoring examples might be constructed in the form of Thurstone-scaled series.

Two main conclusions can be drawn from this study. First, some of the conditions under which distribution-restricted and -unrestricted measurement techniques produce different (and sometimes contradictory) results have been experimentally demonstrated for the field of documentation. This should be useful for helping to select methods of measurement, and it should stimulate thought about measurement problems in this field in general. Second, the results from these and other studies in this series strongly suggest that the area of documentation system phenomena, unlike some other apparently equally complicated social communication phenomena, lends itself especially well to controlled experimentation of the kind reported here. This raises the hope that dogged but continuous progress can be made in reducing the apparent complexity and sometimes capricious quality of document system phenomena, through articulated series of experiments and analysis.

VI. ADDITIONAL WORK IN PROGRESS

The SDC Relevance Assessment Project is continuing to develop additional evidence on many aspects of relevance judgments. One study in progress that is related to the preceding discussion involves the question of the appropriate number of categories for category scale judgments of relevance. Scales with two, four, six and eight categories are being compared, in a population of information specialists and information scientists, for reliability, judgmental certainty, and ease of use.

Another project with similar goals being conducted at Western Reserve University's Center for Documentation and Communication Research by REES, SCHULTZ and others [10] is providing interesting results, leading to the same general conclusions: The complexities of document communication phenomena will not be thoroughly understood, clarified, and resolved in this year or even in the next few years, but controlled analytic experiments are unmistakably capable of increasing our applicable knowledge, significant step after step.

VII. REFERENCES

[1] C. A. CUADRA and R. V. KATTER: Opening the black box of relevance. *J. Docum.* (in press, 1967).

[2] C. A. CUADRA, E. H. HOLMES, R. V. KATTER and E. M. WALLACE: Experimental studies of relevance judgments: Second Progress Report. TM-3068, System Development Corporation, Santa Monica, July 1966.

[3] C. A. CUADRA, E. H. HOLMES, R. V. KATTER and E. M. WALLACE: Experimental studies of relevance judgments: Third Progress Report. TM-3347, System Development Corporation, Santa Monica, January 1967.

[4] W. RIDDLE, T. HORWITZ and R. DIETZ: Relevance feedback in information retrieval systems. Section VI in: Information Storage and Retrieval. Scientific Report No. ISR-11 to the National Science Foundation. Department of Computer Science, Cornell University, Ithaca, New York, 1966.

[5] V. E. GIULIANO and P. E. JONES: Study and test of a methodology for laboratory evaluation of message retrieval systems. Interim Report. ESD-TR-66-405, Decision Sciences Laboratory, Electronic Systems Division, Air Force Systems Command, United States Air Force, L. G. Hanscom Field, Bedford, Mass., August 1966.

[6] H. E. STILES: The association factor in information retrieval. *J. Assoc. Computing Machines*, 1966, **8** [2], 271–279.

[7] C. CLEVERDON, J. MILLS and M. KEEN: *Factors Determining the Performance of Indexing Systems.* Volume 1. Design. ASLIB—Cranfield Research Project, England, 1966.

[8] S. S. STEVENS: A metric for the social consensus. *Science N.Y.*, 1966, **151**, 530–541.

[9] S. S. STEVENS: On the operation known as judgment. *Amer. Scientist*, 1966, **954**, 385–401.

[10] A. M. REES and D. G. SCHULTZ: A field experimental approach to the study of relevance assessments in relation to document searching. Formal Progress Report No. 3, NSF Contract No. C-423. Center of Documentation and Communication Research, School of Library Science, Western Reserve Univ., Cleveland, Ohio.