The Simple Scalability of Documents

Mark E. Rorvig
Johnson Space Center, JL, National Aeronautics and Space Administration, Houston, TX 77058

The relationship between scaling practice and scaling theory remains a controversial problem in Information Retrieval research and experimentation. This article reports a test of a general theory of scaling, i.e., Simple Scalability, applied to the stimulus domain of documents represented as abstracts. The significance of Simple Scalability is that it implies three important properties of scales: transitivity, substitutibility, and independence. The test results indicate that, with some reservations, this theory of scaling is applicable to documents. This finding is further applied to the construction of test collections for Information Retrieval research that could more sensitively measure retrieval system alterations through the use of documents scaled not merely by relevance, but rather, by preference.

Introduction

Work in Information Retrieval (IR) in the early 1960s assumed that preferences for documents were both stable and reliable. It was assumed that it was only necessary to "measure" the "performance" of a system in terms of its ability to deliver documents according to "known" criteria. It was not until the late 1960s that researchers focused on the measurement of human judgments in system evaluation studies rather than other factors such as indexing languages and mechanical system features. However, this focus neglected much of the substantive theoretical work on human judgment in other fields. This absence of relations between scaling practices used for such measurements and scaling theory yielded results of great metric indeterminancy (Caudra, Katter, & Holmes, 1967; Rees & Schultz, 1967). Only one researcher of this period suspected that these findings might arise from the absence of an underlying set of theories (Katter, 1968).

This article reports an experiment that makes an explicit test of a general theory of scaling applied to the domain of documents. This theory, called Simple Scalability, arises from two separate models of human judgment in Mathematical Psychology: The Law of Comparative Judgment developed by Thurstone (1927a, b), and the Choice Axiom of Luce (1959). Although the similarity between the two models was recognized early on (Coombs, 1964, p. 369), it was not until 1972 that both models were unified into the more general notion of Simple Scalability by Tversky and Russo (1972).

The tenets of Simple Scalability, if accepted for a given class of stimuli, suggest the possibility of calculating the probability that a member of a set will be chosen from the set on the basis of its assigned utility, or preference scale value. Consider a set with only three members, x, y, and z with a unique utility U assigned to each member. If the members of this set are simply scalable, then there exists a law of correspondence between the set members and the values of their assigned utilities such that if x should possess a higher value U than y and y a higher value U than z, x will be preferred to z or y no matter whether z is compared to x or y is compared to x; that is, the utilities are transitive. Moreover, such utilities may be thought of as parameters that may be estimated from the random variable reflected in the ratio of the choice of one member of a set divided by the number of times each alternative member in the set was presented. Thus, two set members which themselves differ may have the same parameter value U (or, in this discussion, the same preference scale value), and may be substituted for one another without affecting the scale values of the other set members. Finally, withdrawal of any set member or members will not affect the utility values of the members remaining in the set, that is, the utility parameter of each member exists independently of the composition of the choice set. (These ideas are discussed in more detail in Appendix A of this article.)

The essential question for IR posed by Simple Scalability is this one: From observed choice behavior of human beings for documents, can utility parameters be estimated for documents within the theoretical stipulations of the theory? If Simple Scalability is true for documents, then transitive orderings of documents may be obtained from any convenient subset of the collection. Moreover, these utilities will be independent values, unaffected by documents that the searcher does not review.

Received March 21, 1987; revised October 23, 1987; February 24, 1989; accepted March 1, 1989.

The value of these capabilities for the field of IR has been noted since the 1970s as formally defined by Cooper (1973) and Kochen (1974). In sum, if the postulates of Simple Scalability are accepted for document judgments, then it would be possible to discard the notion of "relevance" in system evaluation and substitute the measurement of document preferences. This is important because of the difficulty posed in system evaluations by binary judgments of relevance or by judgments of relevance obtained by subjective estimate methods, (since such methods, usually ratings in IR experimentation, are known to create distortions in judgments rendered on complex stimuli (Rorvig, 1988)). For example, consider 10 documents retrieved in a given search where all 10 are judged relevant to a stated information need, while only a few possess high utility. Without a measure providing utilities on an interval or ratio scale for individual documents in a retrieved set, the performance of two systems each retrieving 10 relevant documents could not be compared; their system value would be approximately equal. Measures of preference stronger than the ordinal level ones derived from subjective estimation or binary judgments are needed to evaluate alternative systems and alternative system designs with greater accuracy and sensitivity. The weak measures presently associated with these tasks imply a commensurately weak ability to control the various aspects of the IR process.

Additionally, new literature both summarizing the general topic of scaling and evaluating specific aspects of scaling has recently appeared (Rorvig, 1988; Eisenberg & Barry, 1988). This article should be viewed as a further contribution to this emerging class of studies. At the conclusion of this article, an application of scaling techniques flowing from the theoretical propositions of Simple Scalability is discussed; specifically in relation to the problem of constructing test collections for IR based on interval scales of document preferences rather than relevance judgments obtained by ordinal level subjective estimation or binary judgment methods.

Experimental Design

A test of Simple Scalability requires at least two separate tests; one for transitivity and the other for independence. A third test for substitutibility would be useful, but is not required in the context of this study. In structure, the two test requirement boils down to staging an experiment within an experiment. This structure also has the added advantage of permitting some contrasts between two methods for scaling human judgments: Comparative judgments using a large number of judgments under the assumptions of Thurstone's Law of Comparative Judgment; and subjective estimates obtained by a rating scale under the assumptions of Luce's Choice Axiom. A third scaling method, the method of constant sums, used in the earlier study of Rees and Schultz (1967) is also discussed for contrast purposes, though it is not essential to the Simple Scalability test itself.

For the transitivity portion of the test, 51 subjects, drawn from a large pool of business school upper-division undergraduates, were exposed to a business problem concerning self-regulation in advertising (Aaker & Myers, 1981, pp. 537–538) according to a methodology employed by Rees and Schultz (1967). (The key phrases of the text of instructions to the subjects are reproduced in Appendix B.) Subjects then viewed 10 abstracts of documents in 45 pairs, ordered randomly according to the methodology prescribed by Ross (1934). The document abstracts were retrieved under typical online search conditions from the database ABI/Inform and selected randomly from the retrieved set. For each pair of abstracts subjects were instructed to record the member of the pair that would be most useful to them in conducting research on the prescribed business problem. In all, 2,385 judgments were collected.

For the independence portion of the test, subjects were randomly assigned to three groups. The first group of 30 subjects viewed an alphabetized list of all 10 document titles paired with a single abstract. The pairs of document lists and abstracts were also randomly ordered following Ross (1934). Subjects in this group viewed each abstract twice and recorded their judgments on a 10-point subjective estimation scale, assigning the scale values according to the degree that each abstract from all 10 abstracts would be useful in conducting research on the prescribed topic. In all, 600 judgments were collected from this group. The second group of 12 subjects received the same treatment as the first group, expect that they were shown only the first five document list/abstract pairs from the alphabetized list and provided responses four times to each abstract. The third group of 11 subjects viewed the last five document list/abstract pairs from the alphabetized list. Groups two and three provided 240 and 220 judgments respectively. All groups completed the second experimental procedure immediately upon completion of the first, comparative judgment procedure. Although no individual times were recorded, the first experimental procedure required about 50 minutes for all subjects to complete. The second procedure required about 20 minutes for completion by all subjects.

Regarding the transitivity component of the test of Simple Scalability, the reader will recall from the earlier discussion that the utility value, or probability of choice, of a member of the set of 10 abstracts may be estimated from the ratio of choices of each member of the set divided by the number of times the choice was offered. The transitivity test is therefore only a test of the validity of these estimated utility parameters represented by the matrix of ratios of choices to rejections. Two statistical tests were used: Mosteller's Least Squares Test (Mosteller, 1951a, b) and Kruskal's MONANOVA (Kruskal, 1964, 1965; Kruskal & Carmone, 1968). The Least Squares Test determines whether the matrix of choice ratios differs significantly from a theoretical matrix derived from the normal probability distribution. Significant differences are reported by chi-square scores. MONANOVA uses multidimensional scaling to calculate a measure of "stress" for the matrix where stress represents the "goodness of fit" to the additivity hypothesis required for transitivity to hold for the matrix values. The intuitive interpretation of stress scores rendered by Kruskal (1964,

p. 2) is that scores below 15% represent a "fair" fit while scores above 20% and below 5% represent poor and good fits respectively. The two tests together represent complementary methods of evaluating the transitivity of the utility parameters obtained empirically by the Law of Comparative Judgment.

The independence component of the Simple Scalability test was examined by comparing the ratings of the full and two half sets of abstracts. The assumption in this test is that, if the subjects had indeed provided a valid scale by the pair comparison method employed in the transitivity test, then the scale derived from their subjective estimates by the rating method should correspond strongly to the pair comparison derived scale. Further, the rating scales provided by both groups on the half sets should also correspond, when combined, to the rating scale provided for the entire set of ten abstracts. Specifically, the half set orderings should not be affected by the withdrawal of the other half of the abstracts, that is, the values should be independent of the withdrawal of the other set members as stipulated in the earlier discussion of the preceding section. Association among all sets of scales reduced to their column means was measured by Pearson's R score of product moment correlation.

Results

Four graphs are presented in this section. All graphs represent matrices of choice ratios or cumulative probabilities of rating scores as tested for stress by MONANOVA and rescaled to represent the best fit of the data to the requirements of transitivity. Graphs two and four comprise the main points of interest from the discussion of the prior section, but graphs one and three require some further commentary. The tables of these data are available from the author, and may also be found in Rorvig (1985).

Figure 1 represents data exhibiting bias due to "space error effects," or the unintended consequence of failure to preexpose the subjects sequentially to the randomly ordered

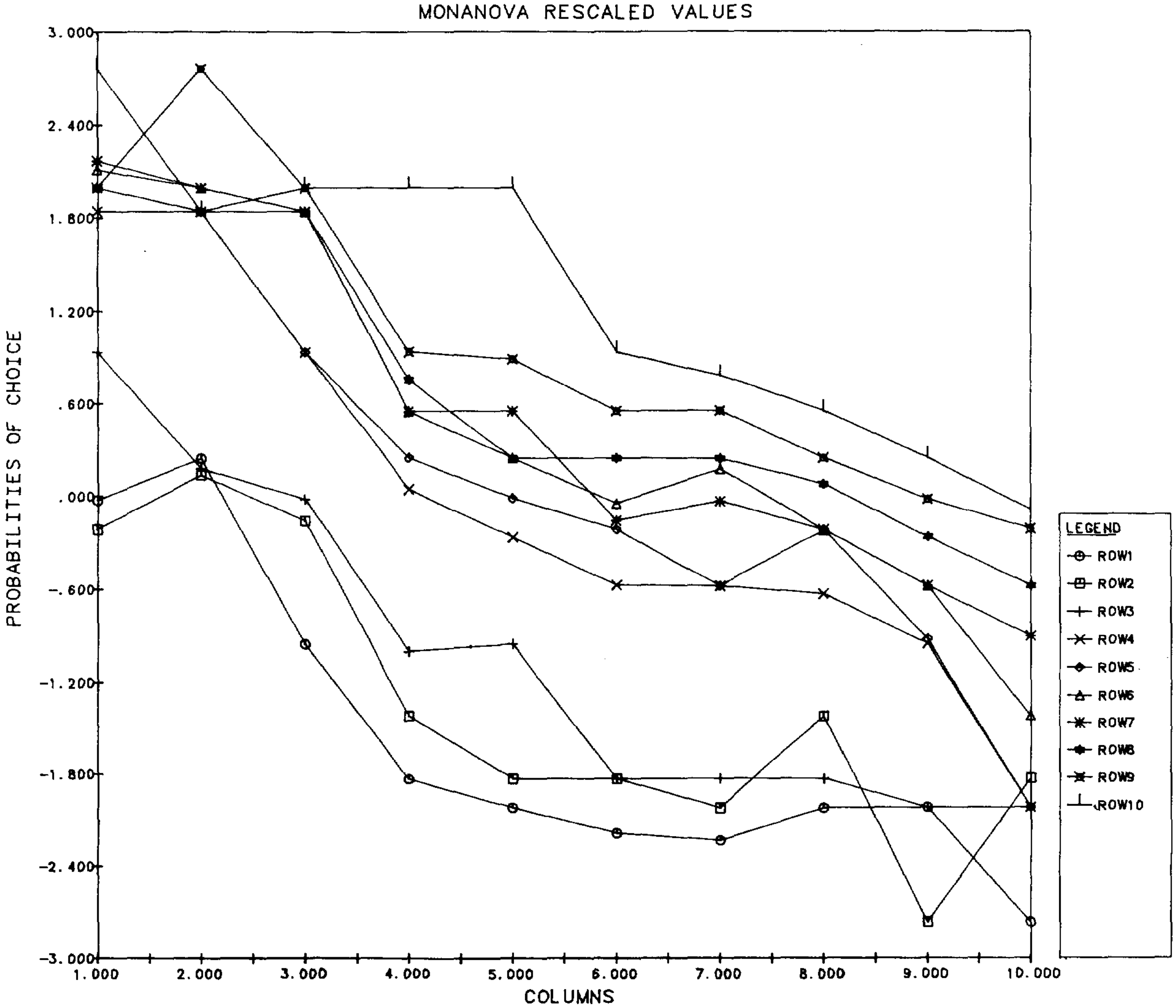

FIG. 1. Initial data set from the transitivity test. The data bias due to order effects is apparent in the upper-right crossovers between row 1 and row 2. Data from column one of the source matrix were removed to produce graph two.

abstracts before asking them to make any pair comparison judgments. Figure 1 nicely displays the order effects discussed by Eisenberg and Barry (1988), but intended in their case. In this case the bias results from presentation to subjects as the first pair, the set members judged most and least useful by them. The unfortunate coincidence of this pair as the first pair presented causes a strong degree of intransitivity to appear in Figure 1 for the first two columns and also results in a marginally acceptable score for stress. The effect is removed simply by dropping the first column of choice ratios from the matrix to create the values appearing in Figure 2.

Figure 3 represents the earlier data from the work of Rees and Schultz (1967) obtained by the method of constant sums. The reader will note that of all four data sets so represented, this matrix displays the greatest degree of orderliness and regularity. This is remarkable, since Rees and Schultz (1967) used only 16 subjects to obtain this data, as opposed to the 51 subjects providing the data for Figures 1 and 2. This suggests that the method of constant sums may be both more efficient and more reliable than the methodology employed in this experiment. In this method, described in Comrey (1950), subjects divide 100 points between each member of a pair to express their choice of the member evoking the strongest response. In the present experiment, subjects made a binary choice between pair members. An additional experiment reported by Rorvig (1987) also confirms the superiority of this method of pair comparison judgment elicitation. The results of the transitivity and independence tests appear in Tables 1 and 2.

TABLE 1.

Data Set	Chi-Square	Stress (%)
Figure 1 (Space error data transitivity procedure)	x^2 = 63.02 D.F. = 36 p < 0.99	20.4
Figure 2 (Error free data transitivity procedure)	x^2 = 36.31 D.F. = 28 p < 0.95	14.6
Figure 3 (Rees and Schultz constant sum method)	x^2 = 38.15 D.F. = 55 p < 0.75	7.5

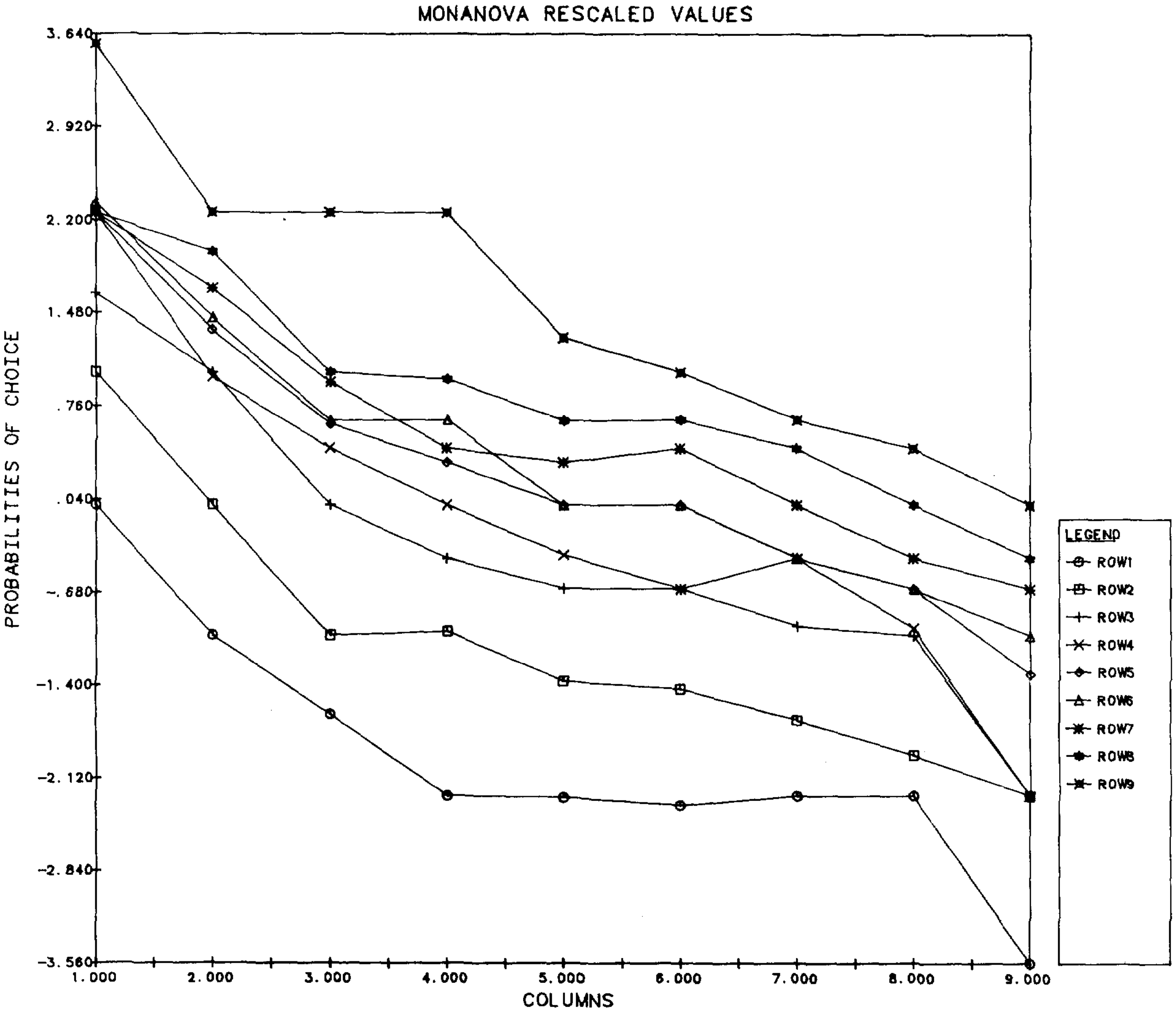

FIG. 2. Error free data from the transitivity test.

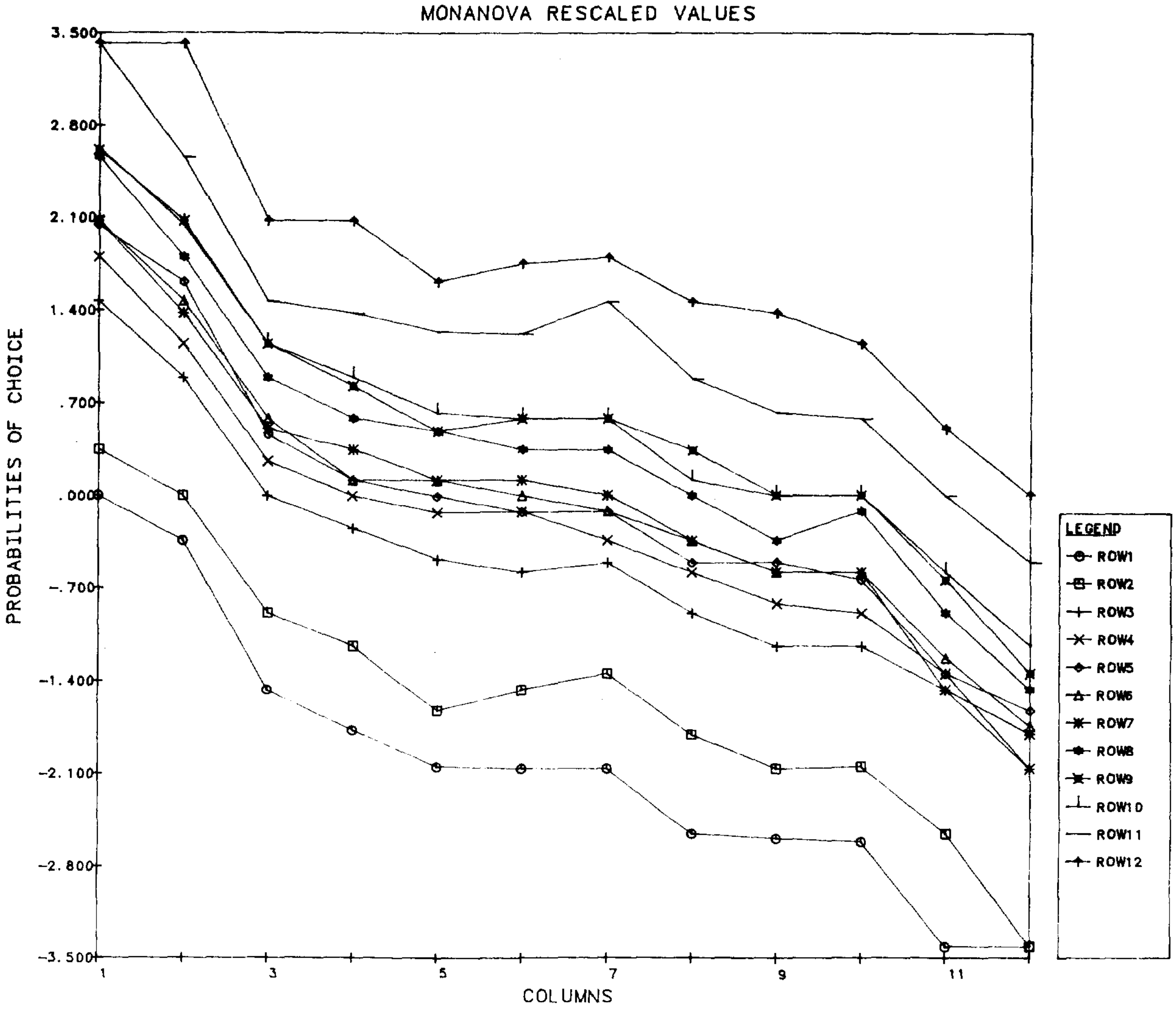

FIG. 3. Data from the Rees and Schultz (1967) investigation at Case Western University. The strong parallelism and absence of crossovers among rows indicate highly transitive document judgments.

TABLE 2.

Data sets	Correlation	Significance
Space error data from Graph One with rating scale data for ten abstracts from Graph Four	$R = 0.9689$	$p = 0.001$
Space error data from Graph One with rating scale data for combined half-sets of five abstracts each	$R = 0.8930$	$p = 0.001$
Rating scale data for ten abstracts from Graph Four with rating scale data for combined half-sets of five abstracts each	$R = 0.9671$	$p = 0.001$

Discussion

With reservations, the experimental results convey a favorable impression for the acceptance of Simple Scalability for the stimulus class of documents as represented by abstracts in this experiment. To a large extent, the conditions set forth for transitivity and independence are satisfied.

Regarding the transitivity hypothesis, the chi-square score for the error free data of Graph Two is well below the 0.95 probability level (set at chi-square = 41.3, with D.F = 28) that was sought, indicating that the empirical data do not differ significantly from the theoretical proportions derived by converting the empirical ratios to unit normal deviates of the normal curve. Moreover, the MONANOVA stress score for this data set of 14.6% is within the "fair" fit range of the test.

However, it must also be noted that these data do not display the high degree of transitivity that would be ideal for Simple Scalability acceptance. An intuitive description of transitivity expressed by Coombs (1964, p. 197) may be helpful in understanding this point. "suppose an individual prefers *A* over *B* 80% of the time and *B* over *C* 70% of the time; then the three levels of stochastic transitivity would require that he prefer *A* over *C* strongly, at least 80% of the time, moderately, at least 70% of the time, and weakly at least 50%." Even weak transitivity by this definition would require the absence of cross-overs in the horizontal lines of graphs one through three, while strong transitivity would require complete parallelism among the lines. Obviously, this did not occur.

In particular, subjects appeared to have difficulty in discriminating among the documents in the midrange of the scale. This phenomena is equally present in graphs one through three, indicating that the effect (at least in these data) is independent of bias conditions or subject composition.

On a happier note, the independence component of the experiment appears to have held rather well, since the correlation between the column means of the half-sets and the full set of 10 abstracts (R = 0.9671) is quite high and significant. The rating scale data for all 10 abstracts as it appears in Figure 4 however, does deserve some additional commentary.

The data in Figure 4 do not represent the typical subjective estimation data taken by rating scales usually found in IR experimentation or test collections. First, these data were taken only after considerable prior exposure by all subjects in the pair comparison procedure. In this procedure, a subject views each abstract nine times, each time in contrast to another abstract in the measurement set. Usual conditions for obtaining ratings data in IR do not include any preexposure that would allow the subjects to derive an implicit scale of the abstracts, where such a scale might serve to make ratings more rigorous by imposing some notion of end points and midpoint as scale anchors. Another difference between these data and typical data is the large number of subjects used (30) and judgments (600) so

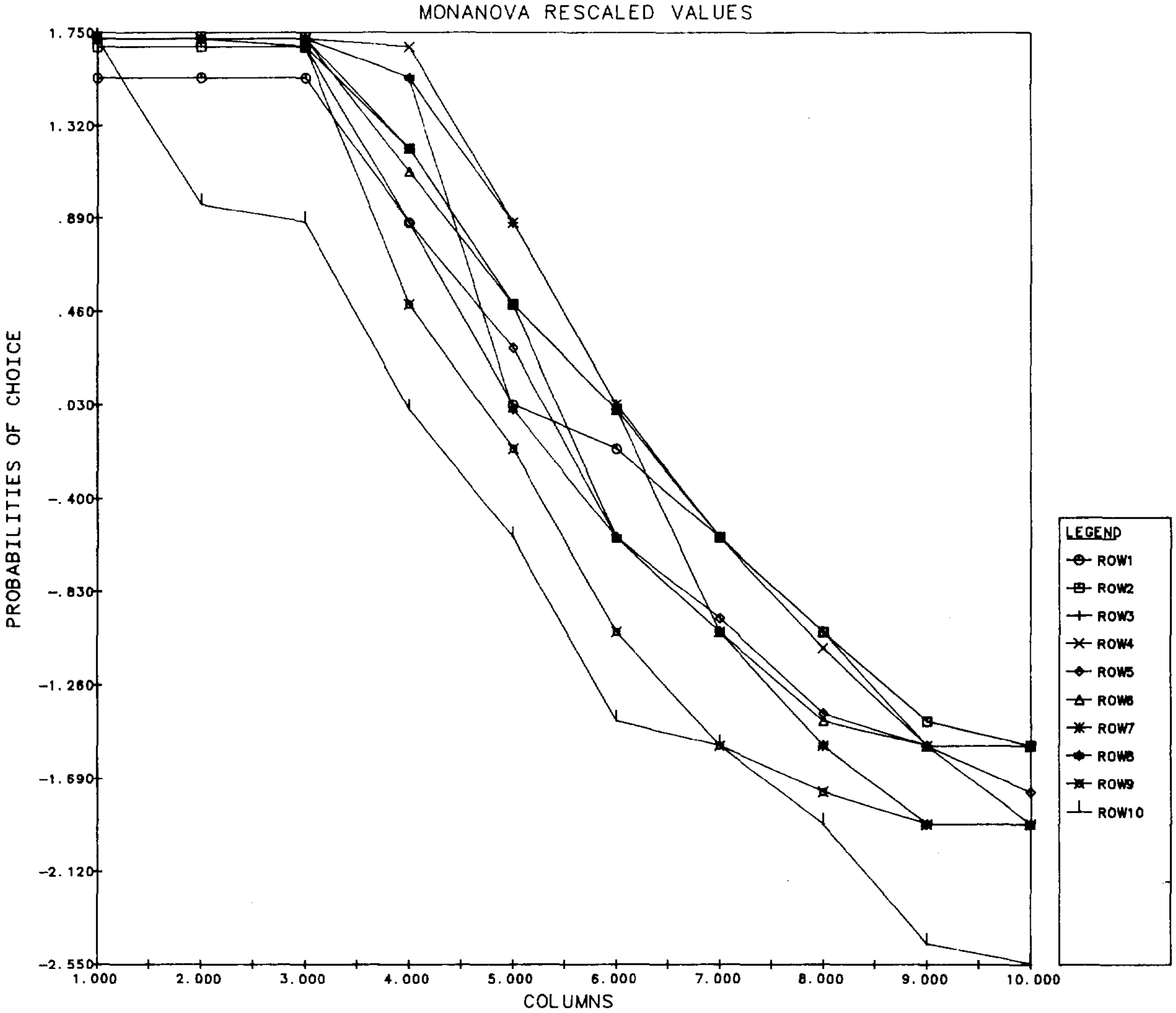

FIG. 4. Data from the independence test obtained by subjective estimation. The absence of parallelism and frequent crossovers among rows indicates imprecision in this method of data collection for document judgments.

obtained for only 10 abstracts. The large number of subjects and judgments would, again, presumably act to make the rating scale more rigorous and orderly.

However, as Figure 4 indicates, none of these factors caused these data to take on the same degree of orderliness and regularity appearing in Figure 2 and 3. It is not so bad in Figures 2 and 3 to have some confusion over abstracts in the middle of the scale, since at least the best and worst document abstracts are clearly delineated. In Figure 4, however, not only are the midscale document abstracts confused, but the end point documents are even more confused! Any attempt to use these judgments for comparative system evaluation purposes would be suspect because it could not even be said with any authority which documents were the worst ones or the best ones for the controlled experimental task.

In a nutshell, this is the predicament of IR research at this time. System designs grow ever more complex in terms of front-end devices, file organization strategies, and statistical analyses of the characteristics of both documents and users. Yet the scale to measure the impact of this growing complexity remains as numb and dull as ever. Though it is freely acknowledged that the data in this article presented in support of the acceptance of Simple Scalability could be stronger, they are a great deal stronger than the usual data used in IR research and experimentation. In the section that follows, Simple Scalability is assumed to be accepted for the stimulus class of documents. The measurement capabilities flowing from this acceptance are then applied to the problem of test collection construction in the form of a proposal for creating test collections reflecting interval scales of preferences.

An Application of Preference Measurement to the Construction of Test Collections

Let us assume that our goal is the creation of a test collection comprising the literature of some socially important area, such as medicine. Let us further assume that the test collection is CD-ROM based, so that its dimensions may be frozen in time. Let us say also that we wish to collect approximately 50 questions to put to this collection. This might be done by establishing contact with a large institution engaged in medical research and teaching. The major stipulation for each question is that a minimum of 15 persons must be homogeneous by the major variables of education, occupation, and research interests, and, willing to judge the retrieved abstracts or documents.

For each question, conduct a search, through an intermediary or any other person, so long as the method is consistent across all questions. Try to obtain as complete an understanding of the questions as possible, with sample references, notes, and taped conversations with the requestors available for the search. Again, the search system and search process does not matter so long as it also is consistent across all questions. Search each question exhaustively, using not only the available system software, but also any other techniques such as author citations, co-citations, and the like, as well as manual search methods. Though this method will be time consuming, it is not far-fetched to conceive it. Moreover it should be noted that exactly this form of seraching was described in an important report on system evaluation (Blair & Maron, 1985).

The resulting retrieved sets may be quite large, with their size dependent on the specificity of the original question. In any attempt to create a test collection of any type, the scarcest commodity is always the time of the judges. Therefore, in the material that follows, it will be suggested that the independence property of Simple Scalability demonstrated in the data of this article plays a major role in the effort to preserve this scarce resource.

From the retrieved set for each question, sample 20 documents or document representations. At convenient times, meet with each member of the group assigned to the specific question. Since 20 stimuli will require 190 pair comparisons, it may be assumed that each individual will require about two to three hours to examine all pairs. The documents may then be scaled by the method of constant sums following the methodologies described or cited in the preceding sections. The scale values resulting from this test will form the basis for scaling the remaining items.

For the sake of illustration, assume that we wish to scale 1000 documents retrieved in response to a question. The initial sample of 20 documents will be used to scale the remaining documents by establishing a brief set of anchors for scaling the remaining abstracts. Since the values of any member of the set may be scaled independently (as indicated by the results in the third section), it will be necessary only to use the best, worst and middle documents to scale all the others. This will result in requiring judges to make, for each additional document, six additional pair comparisons. Assuming that in a typical session, the judge will be given one minute to familiarize herself with the new document, the judgments may be made in less than one minute, since the judge will already be familiar with the other three abstracts. At two minutes apiece, for 980 abstracts, this will require roughly 33 hours of a judge's time and may be spread out over a period of six weeks to two months.

The measure of recall, applied to this test collection, would be approximately the same as it is now for IR test collections. However, the measure of precision would be considerably different, since it could be interpreted over various portions of the abstracts scaled from zero to one. Obviously, the strongest impact of this type of test collection would be on the evaluation of systems that rank documents, since the transitive ranking of the documents by the judges would be quite apparent and would never be a matter of guesswork.

As a final word, let it be noted that this type of test collection would cost a great deal more to build than current collections. However, it may not be a question of whether the field can afford to build this collection, but rather, whether it can afford not to build it. We have entered an era when more documents can be delivered to persons at a lower system cost than ever before; the problem is to retrieve only the most highly useful ones. The theory is at hand to test methods for accomplishing this with greater accuracy than ever before. Something good must surely come of it.

Appendix A: A Formal Definition of Simple Scalability

Notation

Let C be a finite set of some objects, in this case a collection of documents or their representations. Let $A, B, \ldots, N$ denote various nonempty subsets of C, where $\supset$ and $\supseteq$ denote proper and nonproper set membership. The probability of choice of one document from set A shall be indicated $P(x, A)$, and assumed to be estimated by the proportion of choices of x from A over repeated trials. Similarly, let $P(x, y)$ denote the probability that x will be chosen over y. Finally, a scale μ shall be used to indicate a real number function of nonnegative values, that is linear, additive, and unique.

Simple Scalability

Simple Scalability posits the existence of a scale μ defined on members of C such that for any $A = (x, \ldots, z) \supseteq C$

$$P(x,A) = Fn[\mu(x), \ldots, \mu(z)]. \qquad (1)$$

Where Fn is strictly increasing in the first argument and strictly decreasing in the remaining $n - 1$ arguments. In the multi-valued case above, this formula may be rewritten as

$$P(x,A) = \frac{\mu(x)}{\Sigma\,\mu(y)} \qquad (2)$$

or, the Luce Choice Axiom. Moreover, when set A is restricted to two members, equation (1) becomes

$$P(x,A) = F[\mu(x), \mu(y)] \qquad (3)$$

and may also be rewritten as

$$x(\phi) = (\mu(x) - \mu(y)) \qquad (4)$$

the basic Thurstonian Case V model where ϕ denotes the ordering function (Becker, Degroot, & Marschak, 1963, p. 41).

Though quite general, Simple Scalability yields three strong hypotheses for any set of stimuli. First, there is the assumption of strong stochastic transitivity. If, for example,

$$P(x,y) \geq 1/2 \quad \text{and} \quad P(y,z) \geq 1/2 \qquad (5)$$

then, by implication

$$P(x,z) \geq MAX[P(x,y),\ P(y,z)].$$

Second, strong stochastic transitivity implies substitutibility, where, for all x, y, z in C

$$P(x,z) \geq P(y,z) \quad \text{iff}\ P(x,y) \geq 1/2. \qquad (6)$$

Third, both strong stochastic transitivity and substitutibility imply independence. For example, consider a set of binary choice probabilities on x, y, z, w in C. The independence condition is satisfied by

$$P(x,z) \geq P(y,z) \quad \text{iff}\ P(x,w) \geq P(y,w). \qquad (7)$$

Thus, if the choice probability is determined for (x, y) as well as (y, w), then the choice probabilities for (y, z) and (x, w) are also determined. It is of no importance whether $C = (x, z)$ or $C = (y, z, w)$ or any other combination. The ordering function is independent of the composition of the choice set.

Finally, any choice experiment will quickly indicate that at one time a subject may choose $P(x, y)$ and at another time choose $P(y, x)$ where $(x,\ y) \supseteq A \supseteq C$. Thus the proposition from (3) fails unless μ is considered a random vector $(\mu_{(1)}, \ldots, \mu_{(n)})$. In this case, although the subject may be assumed to act rationally by choosing x if $\mu(x) - \mu(y) > 0$ and y when $\mu(x) - \mu(y) < 0$, in practice, the subject cannot do this exactly, and in fact chooses x if $(\mu(x) - \mu(y)) + e > 0$ and y when $(\mu(x) - \mu(y)) + e < 0$. In both the Luce Choice Axiom and Thurstone's Law of Comparative Judgment, e, the error term, is assumed to be independently and identically distributed. (The explicit relation of the two models under this assumption may be found in Luce and Suppes (1965, p. 338) in a proof attributed to A. A. J. Marley.) Under this assumption (3) remains true over repeated trials. Various psychophysical experiments since Fechner reveal this assertion to hold for single subjects only when the stimulus class is very simple (e.g., metal bars of equal size and shape but of different weights.) When stimuli are complex, as is the case for documents, experiments are restricted to data aggregated over individuals.

Appendix B: Instructions to Subjects

Instructions to Subjects for the Pair Comparison Task

Read the two documents carefully. Now remember that I want you to act as if you were actually responsible for investigating the issues of the research statement. For this pair of documents and for each pair of documents in the booklet you have received, choose the one (from the two presented) you would most prefer to have in carrying out research on the issues of the research statement. Indicate your choice by circling the number from the pair of numbers in the recording sheet which corresponds to the document you most prefer to the two documents before you.

Each time you read a pair of documents, please consider only the documents in the immediate pair. Do not think of other documents you have already seen, but concentrate on making an independent choice of one of the two documents before you. Once you have made a choice do not go back and make any changes.

Instructions to Subjects for the Ratings Task

Please turn to the first page of this booklet. On the left hand side of the page you will see a list of ((Groups two and three) some of) the documents you have just seen. The list of documents is in alphabetical order only to help in remembering the documents you have just seen. Now look at the document on the right hand side of the page. You are probably familiar with this document. I would like you to rate this document according to the degree you would prefer to have it relative to the documents listed on the left hand side of the page. For example, if you would prefer this document above all the others in the list, then you would rate it 1, and so on. Note also that each document on the right hand side of the page is numbered.

References

Aaker, D. A., & Myers, J. G. (1981). *Advertising management,* 2nd Edition, Englewood Cliffs, NJ: Prentice Hall.

Becker, G. M., Degroot, M. H., & Marschak, J. (1963). Probabilities of choice among very similar objects: An experiment to decide between two similar models. *Behavioral Science, 8,* 306–311.

Blair, D. C., & Maron, M. E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM, 28,* 289–299.

Comrey, A. L. (1950). A proposed method for absolute ratio scaling. *Psychometrika, 15,* 317–325.

Coombs, C. H. (1964). *A theory of data.* New York: John Wiley.

Cooper, W. S. (1973). On selecting a measure of retrieval effectiveness, Part 1. *Journal of the American Society for Information Science, 24,* 87–100.

Cuadra, C. A., Katter, R. V., & Holmes, E. (1967). *Experimental studies of relevance judgments,* TM-3520/001, 002, 003. 3 vols., Systems Development Corporation, Santa Monica, California.

Katter, R. V. (1968). The influence of scale form on relevance judgments. *Information Storage and Retrieval, 4,* 1–11.

Kochen, M. (1974). *Principles of information retrieval,* Los Angeles: Melville.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29,* 1–27.

Kruskal, J. B. (1964). Analysis of factorial experiments by estimating monotone transformations of the data. *Journal of the Royal Statistical Society, 27,* 251–263.

Kruskal, J. B., & Carmone, F. (1968). Use and theory of MONANOVA, a program to analyze factorial experiments by estimating monotone transformations of the data. unpublished paper, Bell Laboratories, Cherry Hill, New Jersey, 1968.

Luce, R. D. (1959). *Individual choice behavior.* New York: John Wiley.

Luce, D., & Suppes, P. (1965). Preference, utility, and subjective probability. In R. D. Luce, R. R. Bush, and E. Galanter (Eds.), *Handbook of Mathematical Psychology,* vol. III (pp. 249–410). New York: John Wiley.

Mosteller, F. (1951a). Remarks on the method of paired comparisons: I. The least squares solution assuming equal standard deviations and equal correlations, *Psychometrika, 16,* 3–9.

Mosteller, F. (1951b). The effect of an aberrent standard deviation when equal standard deviations and equal correlations are assumed. *Psychometrika, 16,* 203–218.

Rees, A. M., & Schultz, D. G. et al. (1967). *A field experimental approach to the study of relevance assessments in relation to document searching,* 2 Vol. Cleveland, OH: Center for Documentation and Communication Research, School of Library Science, Case Western Reserve University, 1967.

Rorvig, M. E. (1985). *An experiment in human preferences for information in a simulated information system.* unpublished Ph.D. dissertation, School of Library and Information Studies, University of California at Berkeley, Berkeley, 1985.

Rorvig, M. E. (1987). The substitutibility of images for textual description of Archival Materials in an MS-DOS environment. In K. D. Lehman and H. Strohl-Goebel, (Eds.), *The application of micro-computers in information, documentation, and libraries,* (pp. 407–415).

Rorvig, M. E. (1988). Psychometric measurement and information retrieval. In M. E. Williams (Ed.), *Annual Review of Information Science and Technology, 23,* 157–190.

Ross, R. T. (1934). Optimum orders for the presentation of pairs in the method of paired comparisons. *The Journal of Educational Psychology, 25,* 375–383.

Thurstone, L. L. (1927a). A law of comparative judgment. *Psychological Review, 34,* 273–86.

Thurstone, L. L. (1927b). The method of paired comparisons for social values. *Journal of Abnormal and Social Psychology, 21,* 384–400.

Tversky, A., & Russo, J. E. (1969). Substitutibility and similarity in binary choices. *Journal of Mathematical Psychology, 6,* 1–12.