Chapter 4

Discriminating Between Degrees of Low or High Similarity: Implications for Scaling Techniques Using Semantic Judgments*

Renate J. Roske-Hofstrand and Kenneth R. Paap

Several scaling techniques (Pathfinder, multidimensional scaling [MDS], hierarchical clustering) require as input a proximity matrix that specifies the distance between each pair of concepts in the domain of interest. Many applications derive these distance estimates from similarity or relatedness judgments obtained with the method of paired comparisons. For example, ratings have been used to obtain distance estimates between: the display panels of the interface to a flight management system (Roske-Hofstrand & Paap, 1986a), items on a fast food cash register (McDonald, Dayton, & McDonald, 1988), and concepts in an experimental methods course (Goldsmith & Johnson, Chapter 17, this volume). In this method all possible pairs of concepts are presented, one pair at a time, and judges are asked to rate the degree of similarity or relatedness on some scale. This chapter focuses on some issues concerning the sensitivity and reliability of distance estimates derived from ratings.

The primary issue concerns the type of measurement scale inherent in a proximity matrix derived from ratings. On the basis of our own introspections and the debriefing of many raters, we began with the working hypothesis that judges are able to make very fine discriminations between pairs of concepts that are obviously similar, but that it is much more difficult to judge the degree of dissimilarity for a pair of concepts that have *weak* or *secondary* dimensions of similarity.

For example, the reader is invited to rate the similarity of *plate* and *bowl*. Many primary or important dimensions of comparison come readily to mind and most of these comparisons yield good matches, for example, they have the same round shapes (but different depths), they function as holders for food (but different types of food), they are likely to be made from the same material and come in matched sets, and so on. Is the *plate-bowl* pair more or less similar than the pair *plate-cup* or *plate-fork*? To us, these types of judgments seem easy and a feeling of degree of similarity pops up fairly quickly and automatically. This initial feeling can usually be supported by the type of deliberate analysis illustrated above for the dimensions of shape, function, and material.

In contrast, rate the similarity between *washing machine* and *rocking horse*. Sharing our introspections again, if we sense any fast and automatic evaluation to this pair of concepts, it is the feeling that they are completely dissimilar. However, we can search for weak or very abstract dimensions in which they have something in common. For example, the size and structure are such that a child could sit on either one (but is not likely to sit on the former) and movement is an important part of the function of both items. Given agreement that the pair has some small degree of similarity, how does its similarity compare to

*Portions of this study were presented at the 27th Annual Meeting of the Psychonomic Society, November, 1986, New Orleans, LA. We thank Roger Schvaneveldt and Jim McDonald for valuable discussions of this work and Dean Berry and Judy Northrup for collecting much of the data.

sink and *rocking horse*? Our intuitions are that it is much more difficult to discriminate between pairs of concepts with low similarity than pairs of high similarity.

If the working hypothesis is correct, what implications might this have for the way the judges use the rating scale? Two related behaviors seem likely. Assume that judges are asked to rate the similarity of pairs on a scale of 0 to 9 with lower values indicating greater similarity. Given that raters experience many degrees of high similarity they should choose to distribute their responses over several of the values at the similar end of the scale. That is, if the subjective experience is that some pairs are extremely similar, some very similar, some quite similar, and others only moderately similar, then a separate scale value is likely to be used to reflect each of these different states. On the other hand, if degrees of dissimilarity (or extremely low similarity) are difficult to discern, then raters might restrict their responses on the dissimilar end of the scale to only one or two extreme values. This is one prediction we tested in the experiment reported later in this chapter.

If judges are more sensitive to similarity differences among more similar pairs than less similar pairs, then they may also show greater consistency in their ratings. One approach to examining this possibility would be to assess the reliability of judges when they are asked to rate the same set of concepts a second time. This would be a good direct test of the consistency hypothesis, however, we used a somewhat less direct attack that has some other desirable characteristics that will be discussed following a brief presentation of our method.

Suppose an individual has rated the similarity of all pairs in a given domain and is then invited back to the laboratory a few days later. In this second session subjects are presented with two pairs on every trial, such as, *plate-bowl* and *cup-fork*, and asked whether one pair was more similar than the other. Assume for the moment that this judge had, in the first phase of the experiment, rated the first pair as highly similar, say a "1," and the second pair somewhat less similar, say a "2." If the same assessment of similarity occurs during the second task, then the judge should answer "yes," that *plate-bowl* is the more similar pair. However, if the experience of similarity has some instability and *plate-bowl* is assessed to have just a little less similarity during this second session, then the judge may say "no," they seem to have the same degree of similarity. Thus, consistency is reflected in the likelihood that such pairs of pairs are still perceived to have different degrees of similarity.

It should be intuitive that as we present pairs of pairs with greater discrepancy in the original rating that the likelihood of a judge to fail to experience them as different should decline despite some amount of instability. For example, if the pair *plate-bowl* is presented together with a pair that was originally rated as considerably less similar, such as, *bed-dresser*, then *plate-bowl* should still be experienced as more similar, even if, due to instability, it has shifted a bit in the direction of less similarity. The procedure described in more detail in the method section, computes an index of consistency by asking how far down the similarity scale must one go before judges consistently experience a difference in degree of similarity.

The underlying logic is similar to that first used by Weber (1846) in determining the just-noticeable-difference (jnd) between a standard and comparison stimulus at any point along some physical dimension. For example, if we start with the highest pitched tone a listener can hear, how far down the frequency scale would we have to go in order for the listener to consistently report that the two tones differ in pitch? This analogy helps show that our consistency index can also be thought of as a measure of sensitivity. The more sensitive a listener is to frequency differences, then the smaller the difference required to produce consistent "different" responses. Similarly, the more sensitive a judge is to

similarity differences, the smaller the rating-scale difference required to produce consistent "different" responses.

The attentive reader is sure to observe that the psychophysicist who computes a jnd obtains a measure of the subject's sensitivity relative to some physical scale, but that the jnd we compute is relative to the subjective similarity scale used by our judges in the first phase of the experiment. This is true, but psychological sensitivity to rating scale differences can have significant implications for knowledge representations derived from ratings. One such scenario is outlined next.

Suppose we reverse the procedure described earlier and form pairs of pairs by selecting a standard that was originally given the lowest similarity rating. This standard is compared to other pairs which differ by one, two, or three points on the rating scale. How large a discrepancy between the two pairs is required in order for the judges to consistently report that the standard and comparison differ in degree of similarity? We hypothesize that judges are less sensitive to their own rating scale differences when presented with pairs of low-similarity pairs.

If the difference between an extremely similar pair (rating scale = 0) and a very similar pair (scale = 1) is consistently experienced then that one-unit difference in distance in the proximity matrix is psychologically real. In contrast, if the difference between a pair given the lowest similarity rating (scale = 9) and a pair given the next lowest rating (scale = 8) cannot be consistently discriminated, then this one-unit difference is mostly noise. Some scaling techniques will be much more susceptible to this noise than others. For example, MDS tries to consider all pairs of distances simultaneously. In doing so, it gives considerable weight to the longer distances associated with the dissimilar end of the rating scale.

However, Pathfinder is more protected from this noise, since the presence of links is determined, almost exclusively, by differences among highly similar pairs. Dearholt and Schvaneveldt (Chapter 1, this volume) on the foundations of Pathfinder can be consulted for a more complete description of why this is true, but briefly stated the reliance on distances between highly similar pairs comes from Pathfinder's link membership rule. A direct link between two nodes is maintained if, and only if, the weight of that link is shorter than any alternative path that connects those two nodes. Thus, critical decisions determining the structure of the Pathfinder network usually turn on the difference between one short path and another. Links in a Pathfinder network with substantial weights (long distances) are relatively rare and tend to occur as bridges between distinctive subnets or as trails to hermit nodes a great distance from their nearest neighbor.

In summary, our working hypothesis is that judges in a rating task are more sensitive to differences among pairs of highly similar concepts than to differences among pairs that tend to be quite dissimilar. This hypothesis, in turn, leads to the prediction that judges will tend to restrict their responses on the dissimilar end of the scale to only one or two extreme values. It was further suggested that rating scale differences at the high-similarity end of the scale should be more discriminable than equivalent differences drawn from the low-similarity end. Confirmation of the latter prediction would suggest that knowledge representations derived from Pathfinder's scaling of rating data may be less susceptible to bias than those derived from MDS.

Experiment 1

Method

Subjects. Fifty-seven students participated in order to fulfill a requirement of their introductory psychology class.

Procedure for Phase 1 Similarity Ratings. Stimulus presentation was controlled by a Terak 8510 microcomputer. The first part of the displayed instructions were as follows: "The next part of this experiment involves the assignment of similarity ratings to pairs of objects that might be found in a house. Your task will be to judge the similarity or relatedness of several pairs of objects."

The remainder of the instructions described the rating scale and the procedure for making a response:

> In the actual task you will be shown a horizontal rating scale above each pair of terms. The numbers "0" through "9" will be displayed at regular intervals below the line and you are to choose one of these numbers to reflect your judgment. Use "0" for those items that you feel are highly dissimilar and respond with a "9" for items that appear highly similar. The numbers between "0" and "9" should be used to represent degrees of similarity with higher numbers representing greater similarity. Upon responding a vertical bar will move directly above the number you pressed. A second vertical bar will remain at the right end of the scale. The vertical bars also represent similarity. When they are closer together the concepts are more similar.

Two aspects of the instructions deserve discussion. First, the beginning part of the instructions refer to both "similarity" and "relatedness," but the remainder of the instructions refer only to similarity. Furthermore, the rating scale always displayed the phrase "low similarity" at the left end and "high similarity" at the right end. The choice between similarity and relatedness may be important. Some thoughtful individuals have argued that the two are not synonymous and that concepts can be highly related, but not particularly similar. For example, one might judge *can* and *can opener* to be highly related, but not very similar. We doubt that this subtlety was appreciated by most of our subjects, but the effect of similarity versus relatedness remains an empirical issue. If subjects asked the experimenter how to judge similarity, they were told that it was up to them and that there were many different features or attributes that could be considered for any pair of objects.

A second important point to note is that high similarity was represented by larger numbers on the rating scale. This seems more natural for subjects. However, in order to transform these similarity ratings into estimates of distance it was necessary to subtract each rating from nine. Since the distance transformation is required in order to use the ratings as input to a scaling algorithm, all subsequent discussion will refer to the transformed scores. That is, a reference to a 0-pair is a pair of objects that received the highest rating of similarity (a 9 on the similarity scale) and, accordingly, has the shortest psychological distance.

The rating scale with the vertical bar markers described in the instructions was presented in the top half of the screen. The two objects comprising any pair were centered in the bottom half of the screen with one object directly above the other. The object on top was determined randomly. Subjects were presented with all of the 496 possible pairs of

items from their stimulus set of 32 objects. Pairs were presented one at a time, in a different random order for each subject.

Materials for Phase 1 Similarity Ratings. The domain consisted of 64 objects that might be found in a house. The 64 objects were randomly assigned to 4 subsets of 16. Each subject was presented with the items from two of the subsets. Thus, each subject rated the 496 pairs formed from all combinations of 32 objects. Six groups of subjects were required in order to obtain ratings for all pairs in the complete set of 64. The number of subjects in each group ranged from 8 to 11.

Procedure for Phase 2 Discrimination Judgments. One week following their initial ratings, the subjects returned for Phase 2 of the experiment. Stimuli were displayed on a Datamedia 3510 terminal under the control of an APL function that was selecting materials according to the scheme described in the next section. On each trial subjects were presented with a pair of pairs, such as:

1. *spoon-fork*
2. *pillow-cushion*

and were given the prompt: "Is one pair of objects more related than the other?" If the subject pressed "Y" for yes, he was then asked to indicate which of the pairs was more related by pressing either the 1 key or the 2 key.

In retrospect, it may have been better to ask if one pair of objects was more similar than the other since the initial rating task emphasized similarity more than relatedness. It is therefore conceivable that some failures to accurately predict *discrimination in relatedness* in the second phase on the basis of *ratings of similarity* could be attributed to individual pairs having different degrees of perceived similarity and relatedness.

Materials for Phase 2 Discrimination Judgments. To avoid confusion, the following terminology will be adopted. A trial in Phase 1 consisted of the presentation of a pair of objects that the subject rated on the 10-point scale. This pair of objects will be referred to simply as a *pair*. A number in front of a pair indicates its similarity rating (transformed to distance), for example, a 0-pair is a pair of objects perceived to be highly similar. A trial in Phase 2 consisted of the presentation of a pair of pairs (POP) as illustrated above. This type of stimulus will be referred to as a *POP*. The two numbers in front of a POP indicate the similarity rating of each of its constituent pairs, for example, a 0-0 POP is a pair of pairs both of which were assigned the highest degree of similarity.

POPs were selected for each individual subject on the basis of his or her previous ratings of 496 pairs of objects. Four pairs were randomly selected from the set of all pairs given the highest similarity rating (rating value = 0). Five pairs were randomly selected from each of the set of pairs assigned rating values of 1, 2, and 3. Six 0-0 POPs were created from all possible pairings of the four 0-pairs. The four 0-pairs were next crossed with the five 1-pairs, yielding 20 different POPs with a rating-scale difference of 1. The four 0-pairs were similarly crossed with the five 2-pairs and the five 3-pairs. Thus, there were 66 potential POPs drawn from the related end of the rating scale. A full complement of 66 POPs could not be formed for every subject, since not everyone rated a sufficient number of pairs at each of the critical values. For example, if a subject generated only four 1-pairs (rather than 5 or more) then he would be presented with only sixteen 0-1 POPs (rather than 20). Of the 51 subjects who returned and completed phase 2, 29 were presented with the full complement of 66 POPs formed from the related end of the rating scale.

A similar pattern of selections was used at the unrelated end, resulting in full complements of six 9-9 POPs, and 20 each of 9-8 POPs, 9-7 POPs, and 9-6 POPs. Thirty-five subjects received the full complement of POPs at the unrelated end of the rating scale. The set of POPs selected for each subject on the basis of their own earlier ratings were

presented in random order. On any given trial the pair appearing on top and designated as 1 (one) for purposes of responding was determined randomly.

Results & Discussion

Phase 1. Figure 1 shows the distribution of ratings responses (transformed to distance). It is heavily skewed toward judgments of low similarity. As predicted, many pairs (45.1%) are experienced as having an undifferentiated degree of very low similarity.

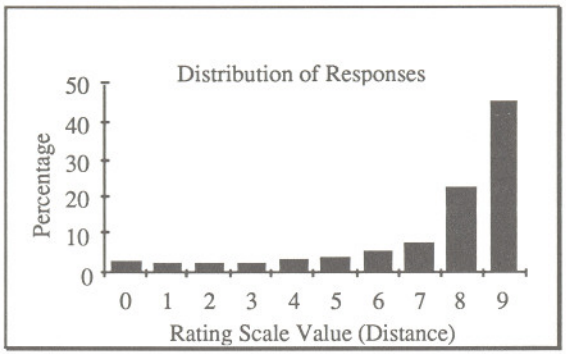

Figure 1. Percentage of each rating-scale response. Low values indicate high degrees of similarity (small psychological distances).

This outcome would be even more pronounced if it were not for an unanticipated characteristic of the response-mapping. Recall that the lowest-similarity value of 9 in Figure 1 is a distance transformation of the actual rating response of 0. Seventeen subjects elected to never use the 0 because "it was out of order on the keyboard." These subjects assigned a 1 to those pairs they judged to be least similar. If this strategy is taken into account by summing the total number of responses in the most extreme category actually used by each subject, the percentage of responses grows to 64.6%. Thus, it appears that almost two-thirds of all pairs are experienced as simply *dissimilar* and assigned the same low value on the rating scale.

There does appear to be a fairly even distribution of responses across the high-similarity values. The percentage of responses across the four values of highest similarity are 2.6, 2.6, 2.5, and 3.2. Although these values cover only 11% of the complete proximity matrix, they roughly correspond to where all the action is during the generation of a *Pathfinder graph.* Path-length differences of 1 or 2 units among these pairs will heavily influence the link structure. Since Pathfinder treats these differences as important, it is informative to determine if people are sensitive to these differences. The results of Phase 2 are, of course, relevant to this question.

Phase 2. The purpose of the second phase was to see how well an individual's original rating could predict his ability to discriminate the similarity of one pair relative to another. POPs composed of two pairs with nearly identical ratings (e.g., 0-1 POPs or 9-8 POPs) should be difficult to discriminate and should receive a relatively low percentage of "different" responses. The percentage of "different" responses should increase as the size of the rating-scale difference increases, for example, the greater similarity of a 0-pair as compared to a 3-pair in a 0-3 POP should be more evident and lead to a higher percentage of "different" responses.

Inspection of Figure 2 shows that the percentage of "different" responses in Phase 2 increases linearly as a function of the size of the rating-scale difference obtained in Phase 1. This is true regardless of whether the standard is the value with lowest or highest similarity. However, the function for differences occurring at the low-similarity end of the scale falls completely below that for the high-similarity end!

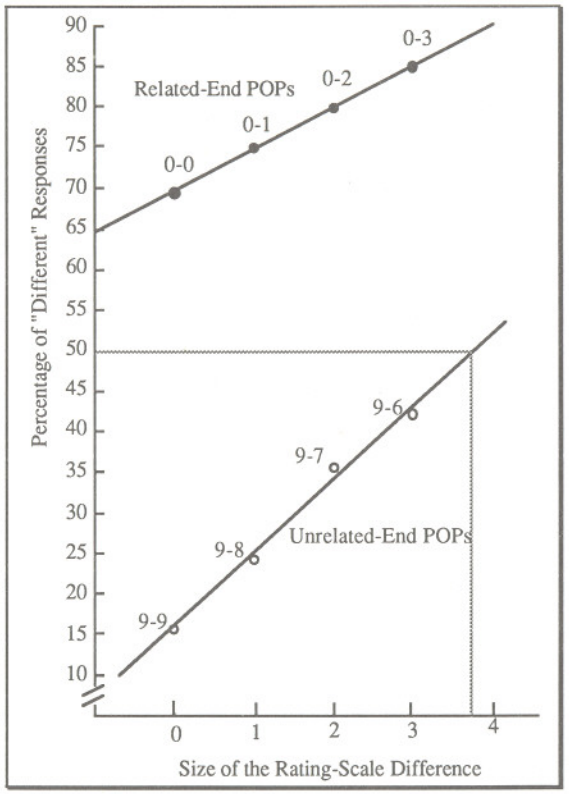

Figure 2. Percentage of "different" responses as a function of the size of the rating-scale difference obtained from the same individual.

The sensitivity index, based on an analogy to the computation of a jnd in classical psychophysics, can be defined as that rating-scale difference required to produce 50% "different" responses. Before examining these results it may be helpful to note that the choice of the 50% criterion cannot be guided by traditional use of the method of constant stimuli. This method usually entails the use of comparison stimuli that are both above and below the standard of interest (Baird & Noma, 1978) and response alternatives of "above" and "below." In this traditional application of the method, 50% "above" responses defines the point of subjective equality, and jnds are usually defined as the stimulus difference corresponding to 25% and 75%. This procedure was not followed in the current study because we were specifically interested in the size of the jnd at the two extreme ends of the

rating scale and, in fact, predicted less sensitivity at the low-similarity end. Having selected the 0-pairs and 9-pairs as standards for theoretical reasons, we could not have comparison stimuli more similar than a 0-pair or less similar than a 9-pair. The criterion of 50% "different" responses is midway between chance and perfect performance under the assumption that subjects would generate 100% "different" responses under conditions of complete discriminability and 0% in the total absence of any sensitivity.

Although, a priori, a criterion of 50% "different" responses seemed reasonable this cutoff cannot be readily applied or interpreted to the data obtained at the related end of the rating scale. Since about 69% of the responses to the 0-0 POPs were "different," we can only infer that the index based on the 50% criterion would be less than one rating-scale unit. Apparently subjects have a tendency to run into the upper end of the rating scale and are capable of making very fine discriminations among highly related items.

The results are quite different at the unrelated end. Even with a rating-scale difference of 3, the function has not reached the 50% cutoff. Linear extrapolation suggests that the size of the index at the low-similarity end of the rating scale is about 3.75 units. Apparently subjects have a difficult time judging the relative degree of similarity between pairs that are only marginally related. Not only do they assign a sizable majority to the lowest-similarity class to begin with, but they also fail to consistently appreciate the difference between this large group of unrelated pairs and those pairs they originally rated as having a little more similarity.

When subjects in Phase 2 indicate that one pair is more related than another, they do not always pick the pair they previously rated as more similar. Figure 3 shows the percentage of "different" responses that were followed by a selection opposite from that predicted on the basis of the similarity ratings. The percentage of reversals is fairly high for the POPs differing by only one unit, but systematically declines with the rating-scale difference. Reversals are indicative of considerable instability in judgments of similarity.

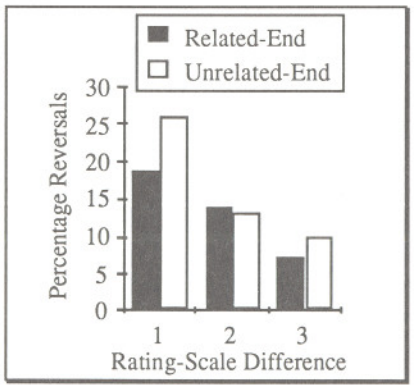

Figure 3. Percentage of "different" responses in which the pair rated less similar during Phase 1 was selected as the more related pair in Phase 2.

It would be informative to determine if the reversals were fairly random or systematic. Rating mistakes in Phase 1 should lead to a systematic pattern of reversals in Phase 2. The following analysis specifically focuses on 0-pairs that may have been overrated and 9-pairs

that may have been underrated. In Phase 2 subjects judged each 0-pair against three other 0-pairs. Based on the original ratings, these should be judged as equally related and receive "same" responses. However, as shown in Figure 2, two-thirds of the 0-0 POPs were judged as "different."

The "different" responses can be used to derive an index of strength for each 0-pair by counting the number of times an individual 0-pair dominates the other three 0-pairs in the corresponding 0-0 POPs. For any given 0-pair, this strength index can range from 0 to 3, with 3 indicating greatest strength. The 0-pairs with the least strength (index = 0) should be most susceptible to reversals. For each 0-pair we determined the number of reversals it produced in all of the 0-1, 0-2, and 0-3 POPs to which it belonged. The strength index was also calculated for each 9-pair based on the number of times it dominated the three other 9-pairs in a subject's 9-9 POPs. In contrast to the 0-pairs, it is the strong 9-pairs that should lead to a greater number of reversals.

Figure 4 shows the mean number of reversals associated with the strength index. Reversals decrease with the strength of a 0-pair and increase with the strength of a 9-pair. It is evident that it is extremely rare for the best 0-pairs to be perceived as less related than a pair previously assigned a lower similarity rating, but that there are some weak 0-pairs that are involved in many reversals. The leading cause of reversals is a strong (index = 3) 9-pair.

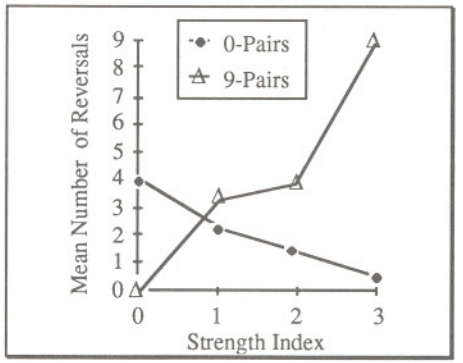

Figure 4. Mean number of reversals (out of 15 possible) as a function of the strength index of the 0-pairs and the 9-pairs.

The left side of Table 1 shows some examples of weak 0-pairs that generated a large number of reversals, while the right side shows examples of strong 9-pairs that produced many reversals. Both the index and the number of reversals for each pair are with respect to some particular individual. Accordingly, given the rules for selecting POPs for each subject the maximum possible number of reversals is 15. The systematic relationship between the strength index and the number of reversals is consistent with the view that some pairs suffer a meaningful shift in similarity from one test session to the next, but that judgments made during Phase 2 are internally consistent. For example, the subject that rated *blender* and *can opener* as completely dissimilar in Phase 1, is clearly treating this pair as having a fair amount of relatedness in Phase 2 and, accordingly, it is selected as the more related pair in 11 of the 15 opportunities in which it was tested with pairs originally rated as 1, 2, or 3 scale units more similar.

Table 1. Examples of weak 0-pairs and strong 9-pairs taken from individual ratings.

Weak 0-Pairs	*Index*	*Reversals*	*Strong 9-Pairs*	*Index*	*Reversals*
desk - night table	0	9	toy chest - mirror	2	9
radio - record	0	4	spoon - record	3	11
book - record	0	7	can opener - blender	3	10
towel rack - hanger	0	4	toy chest - bowl	3	10

Most of the pairs associated with high reversal rates are probably simply mistakes, in the sense that if the judge had it to do over again that pair would be given a different similarity rating. However, some could be caused by the switch in focus from similarity in Phase 1 to relatedness in Phase 2. The *blender-can opener* pair could be viewed as such an example, if this subject judges the pair to be reasonably related, but not particularly similar. Note, that this potential cause does not apply as readily to reversals at the related end. The problem with the 0-pair *book-record* for some particular subject was that this pair is experienced in Phase 2 as having significantly less relatedness than its rating would indicate. The hypothesis that the pair is judged to be highly related, but not so similar, would predict the opposite pattern.

In summary, subjects are more likely to respond "same" to POPs whose original ratings were actually different, if the POP is drawn from the low-similarity end of the rating scale. This is consistent with the hypothesis that it is difficult to discriminate between low levels of similarity. It was also determined that the leading cause of reversals is strong 9-pairs that are judged to be more related than the three other 9-pairs selected for testing. Thus, the low-similarity end of the rating scale is plagued by less sensitivity and greater likelihood of systematic error.

Experiment 2

When ratings are obtained using the method of paired comparisons, scaling solutions can be derived for individual subjects or for an entire group of subjects who have comparable levels of domain expertise. There are potential tradeoffs for either theoretical or applied work. Only individual networks can reflect idiosyncratic differences in conceptual organization between subjects. On the other hand, as revealed above, individual ratings may be prone to *mistakes* and some of the error is likely to be corrected by averaging ratings across subjects. To investigate this issue the second phase of Experiment 1 was replicated with a new set of subjects. In this case the size of the rating-scale difference for each POP was based on the first group's average ratings.

Method

Subjects. Twenty-six judges from the same subject pool participated in the second experiment.

Materials. The POPs for Experiment 2 were selected using the following procedure. The six groups that participated in the first experiment rated a total of 2,016 pairs. These correspond to all possible pairs formed from the complete set of 64 household objects. The mean rating for each pair was rank ordered and the number of pairs within a half unit of each rating-scale value was determined. For example, if the mean for the pair *teddy bear-pillow* was 3.62, the frequency for the rating-scale value 4 was incremented since it was within the range of 3.5 to 4.5. The distribution of pairs within the specified range is shown in the first row of Table 2. Because only two pairs have means less than 0.5, the most related pairs were selected from those within a half unit of a rating-scale value of 1. Nine pairs were selected from the pool of 14 in an effort to both minimize the variance and come as close as possible to a mean of 1.00. The selected means ranged from 0.71 to 1.30, with a mean of 1.00 and a standard deviation of 0.23. Similar procedures were used to select nine pairs with mean ratings close to 2, 3, 4, 6, 7, 8, and 9. These means are shown in the second row of Table 2.

Table 2. Distribution of mean-rating values from Experiment 1 and means of pairs used in Experiment 2.

	Rating-Scale Value									
	0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Number of pairs within a half unit	2	14	29	31	64	101	186	380	911	290
Mean of nine pairs selected for Exp. 2		1.00	2.02	3.00	3.99		6.00	7.00	8.00	9.00

The nine pairs selected for each of the mean rating-scale values were randomly divided into two sets. Set A consisted of five 1-pairs (most similar), five 9-pairs (least similar), and four pairs for each of the intermediate rating-scale values. This permitted the formation of 10 different POPs for rating-scale differences of 0 (i.e., 1-1 POPs and 9-9 POPs), and 20 different POPs for rating scale differences of 1, 2, and 3. Fourteen subjects were presented with the 140 POPs formed from the Set A materials. Set B consisted of four 1-pairs, four 9-pairs, and five pairs for each of the intermediate values. This permitted the formation of 6 different 1-1 POPs and 9-9 POPs, and 20 different POPs for each of the non-zero rating-scale differences. Twelve subjects were presented with the 132 POPs formed from the Set B materials.

Procedure. The subject's task was the same as that used in Phase 2 of Experiment 1. On each trial the subject was presented with a POP and asked to indicate whether the two pairs were the same or different in terms of their degree of relatedness. Whenever subjects responded "different," they were then required to indicate which pair was more related.

Results & Discussion

The same pattern of results was obtained with each set of materials and the analyses described below are collapsed across both sets. Figure 5 shows the percentage of "different" responses as a function of the size of the rating-scale difference for Experiment 2. The results for individual-based (Figure 2) and group-based (Figure 5) stimuli are very similar at the related end. Both would yield a cognitive jnd value less than zero. At the unrelated end, the group-based pairs yield greater discriminability. The 50% cutoff yields a sensitivity index for the group ratings of about 1.75 units compared to the 3.75 when the POPs were formed on the basis of an individual's own ratings.

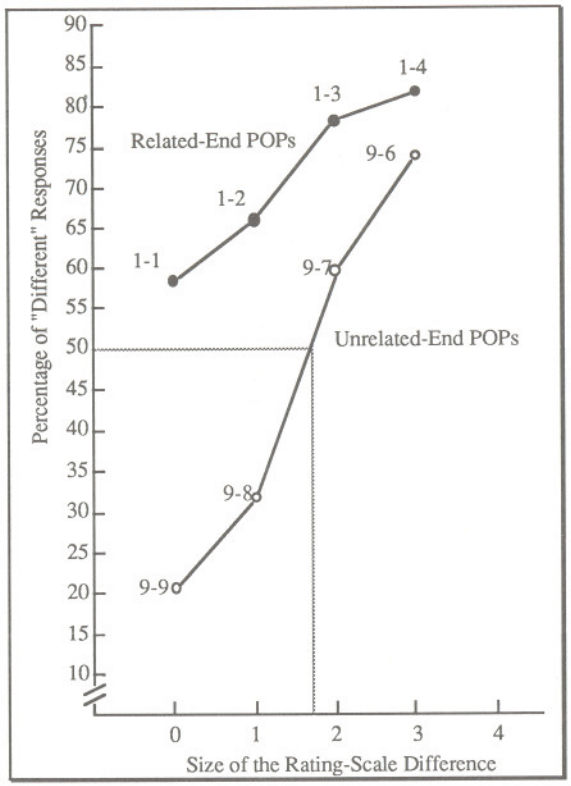

Figure 5. Percentage of "different" responses as a function of the size of the rating-scale difference for the mean ratings from Experiment 1.

Figure 6 shows that the percentage of reversals is associated with the various rating-scale differences. A reversal occurs when the subject responds "different" and then selects as more related, the pair that was rated less similar by the subjects in Experiment 1. The group-based reversals, like the individual-based reversals shown in Figure 3, systematically decrease with the size of the rating-scale difference.

In comparing the percentage of reversals computed by reference to an individual's ratings (Figure 3) to those based on group averages (Figure 6), it appears that the percentage of group-based reversals is somewhat higher at the related end and somewhat lower at the unrelated end. In terms of the tradeoff discussed earlier, this trend suggests that subjects are not in complete agreement when making the fine discriminations among related pairs. However, ratings based on group averages will remove some of the error associated with the difficult discriminations at the unrelated end of the rating scale.

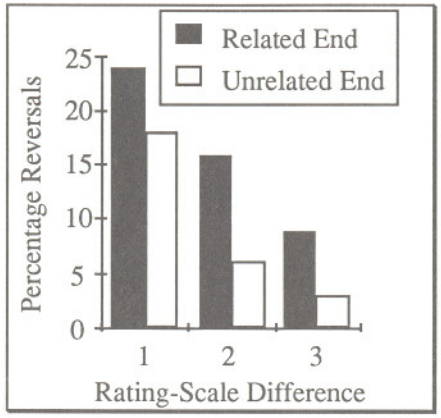

Figure 6. Percentage of "different" responses in which subjects selected the pair rated less similar in Experiment 1 as the more related pair of the POP.

Conclusions

A number of implications follow from these analyses. First, it appears that judges make very fine distinctions among pairs of highly related concepts, but that discriminability suffers at the unrelated end. This is precisely the environment where Pathfinder thrives, since the presence of links is determined, almost exclusively, by differences among highly related pairs. In contrast, scaling techniques like MDS try to consider all distances simultaneously. In doing so, they may be more susceptible to the lack of sensitivity associated with the unrelated end of the rating scale. A second implication is that the conceptual models derived from a homogeneous group of subjects may provide as accurate a model as representations based on an individual's own data. This seems reasonable given that discriminations between degrees of psychological distance could be predicted as well, or better, by the means of an entire group compared to the predictions based on each individual's own ratings. This claim would benefit, of course, from empirical support in which performance was measured as a function of group-based versus individual-based organizations.