

## METHODS & DESIGNS

# An efficient method for obtaining similarity data

ROBERT GOLDSTONE

Indiana University, Bloomington, Indiana

Measurements of similarity have typically been obtained through the use of rating, sorting, and perceptual confusion tasks. In the present paper, a new method for measuring similarity is described, in which subjects rearrange items so that their proximity on a computer screen is proportional to their similarity. This method provides very efficient data collection. If a display has  $n$  objects, then, after subjects have rearranged the objects (requiring slightly more than  $n$  movements),  $n(n-1)/2$  pairwise similarities can be recorded. As long as the constraints imposed by two-dimensional space are not too different from those intrinsic to psychological similarity, the technique appears to offer an efficient, user-friendly, and intuitive process for measuring psychological similarity.

William James (1890/1950) contended that a “sense of sameness is the very keel and backbone of our thinking” (p. 459), and modern psychological research seems to have taken these words to heart. A notion of similarity plays a pivotal role in much theorizing in psychology (Goldstone, 1994; Goldstone & Medin, 1994; Medin, Goldstone, & Gentner, 1993). Some researchers have focused on how similarity assessments are made (e.g., Tversky, 1977), and many more have explored the influence of similarity on other variables. Theories of interpersonal attraction, memory retrieval, concept formation, problem solving, and disease diagnosis, to take a few random examples, are frequently based on similarity (similarity of people, memories, objects, problems, and diseases, respectively).

This paper reports on a new technique for collecting similarity data from human subjects efficiently. The technique involves having subjects arrange items on a computer screen spatially so that similar items are close to each other and dissimilar items are far away. As such, the technique can be thought of as requiring subjects to create their own personal multidimensional scaling (MDS) solutions. In traditional MDS techniques (e.g., those found in Torgerson, 1965), a computer routine is presented with all the pairwise similarities between a set of  $n$  objects. The routine then creates a geometric representation with a point representing each of the objects and the distances between the points representing the

psychological distances between the objects. Objects are arranged so that the proximities between objects approximate their similarities to as close a degree as possible. In the present method, subjects themselves create the two-dimensional geometric representation that best approximates their intuitions about object similarities.

### CURRENT TECHNIQUES FOR COLLECTING SIMILARITY DATA

Similarity data that might serve as input to an MDS routine are typically collected by one of three methods.

#### Ratings

The most straightforward way to collect interobject similarities is to have subjects assign numeric ratings to pairs of items. For example, a subject might be told to assign numbers between 1 and 20 to several pairs of objects, with 1 signifying *very low similarity* and 20 signifying *very high similarity*.

#### Confusions

The similarity between two items is assumed to be proportional to the probability of confusing one with the other. In one paradigm, an item is briefly flashed to the subjects, who are asked to identify it. Sometimes the item will be incorrectly identified as another item. As the similarity of the two items increases, the probability of such a confusion is assumed to increase (Townsend, 1971). In a *same/different* judgment task (see, e.g., Podgorny & Garner, 1979), subjects are presented with two items and are required to give a speeded response as to whether the items are the same or different. As the similarity between two different items increases, subjects are assumed to take longer to respond “different” and to make more incorrect “same” responses.

I thank Herman Gollwitzer, Douglas Medin, Robert Nosofsky, and Richard Shiffrin for many useful comments. The research was supported by a Biomedical Research Grant from the National Institute of Health (PHS S07RR7031N) and by National Science Foundation Grant SBR-9409232. Correspondence should be addressed to R. Goldstone, Psychology Department, Indiana University, Bloomington, IN 47405.

### Sorting

With a sorting measure, subjects are instructed to place items into groups (e.g., Rosenberg & Kim, 1975). It is assumed that the frequency with which two items are placed in the same group is proportional to their similarity.

### Problems With Traditional Measures

Although these methods are valuable tools for investigating similarity, they also suffer from some shortcomings. First, they are fairly inefficient, requiring many judgments before reliable results can be obtained. If one wishes to create an MDS solution for  $n$  objects, one will have to collect at least  $n(n-1)/2$  similarity ratings. The first object must be compared with the second, third, . . . , and  $n$ th object, the second object must be compared with the third, fourth, . . . , and  $n$ th object, and so on. A reliable MDS solution will require even more judgments if a perceptual confusion measure is used. Some objects will be confused with each other to a significant but small degree. If this is the case, many object presentations (upward of 1,000 for each subject in Townsend, 1971) will be required. Because the number of required comparisons increases as a quadratic function of the number of compared objects, very few applications of the confusion or rating measures have been used to compare more than 40 items. Requiring many judgments to be made is problematic, not only because of the lengthy experimental sessions needed, but also because subjects may change their strategies with practice. The similarities that are obtained from these methods may not correspond closely with naive subjects' assessments.

Second, the sorting and confusion techniques do not admit of graded individual responses. For the confusion measure, an object is either confused with another object or not (and an object can only be identified as one other object on any given trial). For the sorting measure, objects are either placed in the same categories or not. Graded estimates of similarity emerge through averaging over binary similarity measurements, but sorting and confusion measures suffer from low resolution on individual trials because graded responses are not admitted (though Townsend, 1971, did collect confidence judgments as well).

Finally, these methods do not take advantage of certain systematicities in similarity data. For example, in many situations, if  $A$  is similar to  $B$ , and  $B$  is similar to  $C$ , then  $A$  will be similar to  $C$ . If  $A$  is similar to  $B$ , and  $B$  is very dissimilar to  $C$ , then  $A$  must be dissimilar to  $C$  too. If these relations apply to a stimulus set, then a measure that intrinsically embodies these systematicities will be able to avoid the redundancy of the other measures.

### SIMILARITY VIA SPATIAL ARRANGEMENT

In the new similarity measure proposed, subjects are given a random configuration of objects, and they re-

arrange the objects so that the distances between them are proportional to their dissimilarity. This technique does not seem to suffer from the disadvantages of the other measures. First, the technique is quite efficient. Whereas 2,016 similarity ratings would be required to compare 64 objects with each other, a rearrangement of the 64 objects on a single screen can hypothetically provide the same information. The technique uses the quadratic complexity of interobject comparisons to its advantage rather than its disadvantage. When an object is moved from one screen location to another, 63 interobject distances are altered. Every time 64 objects are rearranged, 2,016  $[n(n-1)/2]$ , where  $n$  is the number of objects] interobject distances can be recorded. Given the spatial constraints of a two-dimensional field, these 2,016 data points are not independent of each other, but the efficiency of data collection is clear.

Second, graded similarity data are obtained. Similarity is quantified in terms of the screen distance between items, as measured in centimeters or by computer screen pixels.

Third, the currently proposed method embodies reasonable constraints on the incoming similarity data. If items  $X$  and  $Y$  are placed very close to each other, and  $Y$  is placed close to  $Z$ , then  $X$  will have to be placed fairly close to  $Z$  as well. Although metric assumptions for psychological similarity are not always met (Goldstone, Medin, & Gentner, 1991; Tversky, 1977; see also the General Discussion), they provide a useful approximation that eliminates much of the redundancy of typical similarity measurement techniques.

An experiment was conducted in order to assess the efficiency and reliability of the new technique for obtaining similarity data. Results obtained from the procedure are compared with the data collected by traditional similarity ratings and *same/different* confusions.

## METHOD

### Subjects

Thirty-five undergraduates from Indiana University participated in the spatial arrangement procedure, 35 more undergraduates participated in the similarity rating procedure, and 30 undergraduates participated in the *same/different* judgment procedure. No subject participated in more than one condition.

### Materials

Sixty-four uppercase As were obtained from *The Type Specimen Book* (V&M, 1974). The letters were chosen so that they would be readily distinguishable from each other. Figure 1 illustrates several of the letters. Each letter was transferred by an analog-to-digital scanner to Macintosh IIsi computers. The average dimensions for the letters were 2.8 cm high  $\times$  2.1 cm wide.

### Procedure

The subjects in the spatial arrangement group were given the following instructions: "You will be shown a set of letters on the screen. You should move the letters around so that letters that are similar to each other are close. The more similar two letters are, the closer they should be." The subjects were also instructed on how to use a mouse to move the letters. When the mouse-controlled cursor was on top of a letter to be moved, subjects were

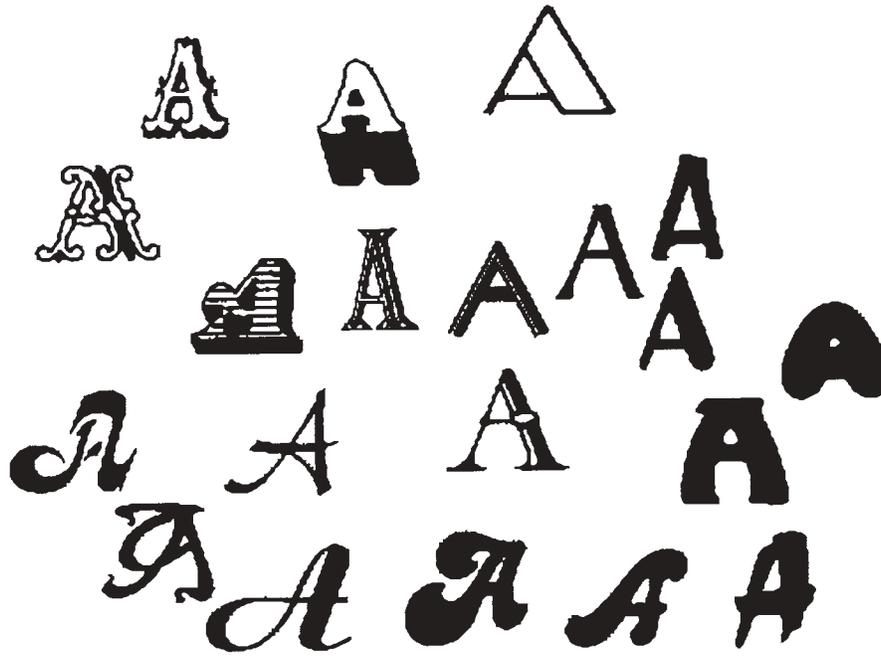


Figure 1. A sample display of letters after a subject has rearranged the letters so that the proximity of any two letters is proportional to their similarity.

instructed to press the button on the mouse. The letter would disappear. The subject would then move the cursor to the new desired location for the letter, and would press the mouse button again. The letter would reappear in the new location. This procedure was continued until the subject was satisfied with the spatial configuration of the letters, whereupon he/she would move the cursor into a box labeled "DONE" and would press the button. The computer then recorded the distances (in pixels) between each pair of letters. The data were obtained and recorded on Macintosh IIsx computers. The screen was 25.4 cm wide  $\times$  19.1 cm high. The letters were displayed on a 22.9 cm  $\times$  17.8 cm region on the screen. The screen was viewed from a distance of approximately 61 cm.

On each display, 20 letters were presented simultaneously. These letters were selected at random from the set of 64 letters. They were arranged in five columns of 4 letters each. The particular location of each letter was randomized. Each subject received 30 displays in all.

The subjects in the similarity rating procedure assigned ratings between 1 and 9 (1 = *not very similar at all*, 9 = *very similar*) to pairs of letters. A subset of 15 of the 64 letters was tested. Each of the 15 letters was compared with each other letter, yielding a block of 105 trials. Compared letters were separated by 10 cm. Each subject received seven such blocks. The order of the letters on each trial and the order of trials were randomized. The subjects were shown 15 sample pairs of letters and were told, "These are the type of comparisons that you will be asked to make. Try to use the entire range of ratings."

The subjects in the *same/different* judgment task received the same pairs of letters as did subjects in the similarity rating task, and they also received an equal number of pairs in which the two letters were identical. The position of the letters on each trial and the order of trials were randomized. The subjects were instructed to press one key to respond that the letters were the same, and another key to respond that the letters were different. They were encouraged to respond as quickly as possible without sacrificing ac-

curacy. Each subject received four blocks of the trials, yielding a total of 840 trials.

## RESULTS

Comparisons will focus on the similarity relations between the 15 letters that were included in all three methods for obtaining similarity data. It should be noted, however, that the spatial arrangement technique obtained complete pairwise data for 64 letters in approximately the same amount of time per subject (1 h) as the other methods required in order to obtain data for only 15 letters. Projections indicate that the other methods would require more than 19 h to collect comparably reliable data for all 64 letters.

There is no test-independent method for revealing "true similarity" (Goldstone & Medin, 1994). As such, one of the only objective ways to assess whether a measure is capturing some notion of similarity is to note its degree of correlation with other measures that are assumed to be partially capturing similarity. The Pearson correlations between the three measures are as follows: correct *different* response times and ratings,  $r = .85$ ; correct *different* response times and spatial arrangement,  $r = .87$ ; ratings and spatial arrangement,  $r = .93$ . The correlation between the rating and spatial arrangement measures is significantly higher (Fisher's  $r$ -to- $Z$  transform,  $p < .05$ ) than the other two correlations, and all correlations are significantly greater than zero.

Subjects in the spatial arrangement task made an average of 24.3 object movements, and spent an average of

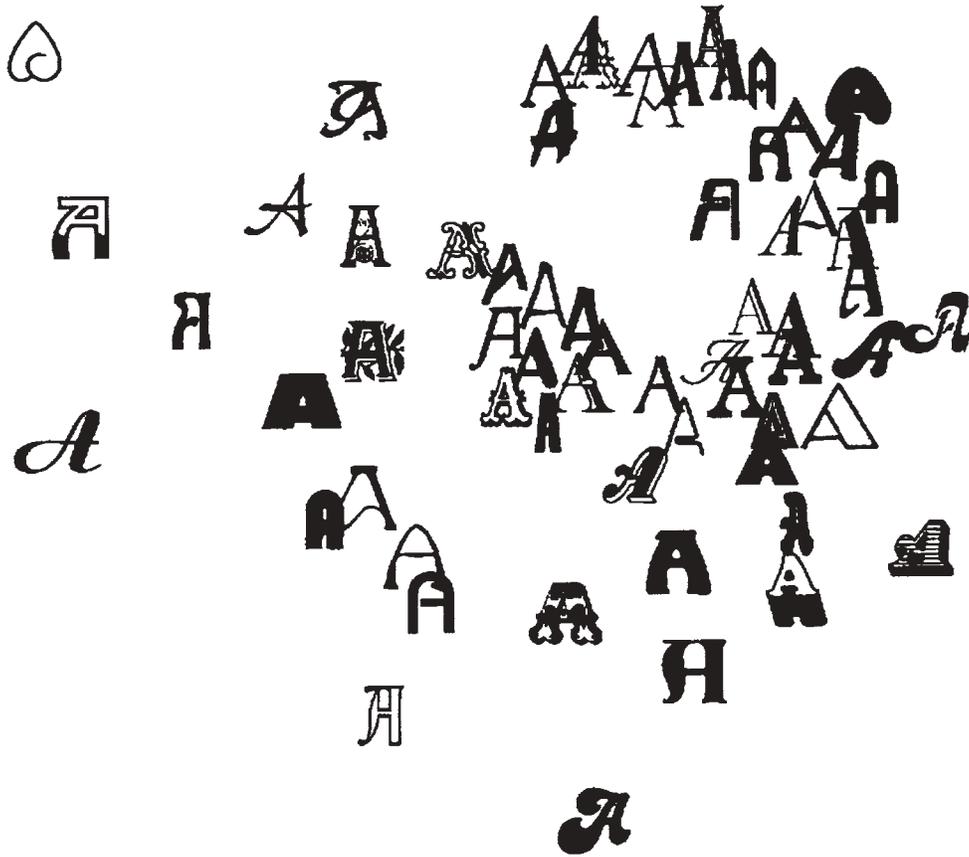


Figure 2. A multidimensional scaling solution of the data obtained from the spatial arrangement technique. Only two of the five dimensional coordinates of the letters are shown.

1.7 min, for each display. An MDS solution was constructed for the full 64-letter set shown to subjects in the spatial arrangement task. Figure 2 shows a plot of the letters along the two most important dimensions from the MDS solution. Even though subjects were limited to two dimensions for their responses to individual displays, MDS solutions with greater than two dimensions reduced the stress of the MDS solution substantially. The fit of the MDS solution significantly increases with each added dimension from two to five. Increasing the dimensionality beyond five does not improve the fit further, as is indicated by small decreases in stress. The stresses for the one-, two-, three-, four-, five-, and six-dimension solutions are 338, 242, 178, 140, 123, and 115, respectively. Thus, the spatial arrangement method can be used to isolate many stimulus dimensions, even though only two dimensions are necessary to account for any individual display's similarity relations.

## DISCUSSION

The results show that the data obtained from similarity ratings are highly correlated with data obtained by the newly described spatial arrangement technique.

Moreover, the two measures are equally good predictors of subjects' performance on a *same/different* judgment task. Thus, the new technique appears to yield data whose reliability is comparable to that of data produced by other traditional techniques.

The technique has several advantages over other methods. With this technique, data are collected much more quickly than with traditional methods. The program uses an intuitive and attractive interface that takes advantage of people's natural tendency to think of similarity spatially. This tendency is suggested by statements such as "these two styles are quite close to each other" and "the distance between their stances is quite wide." The technique allows for graded degrees of similarity rather than binary *similar/dissimilar* judgments. Finally, the technique places each similarity decision in a context of other decisions; with ratings, the context emerges only with time (the rating assigned to the first pair of objects presented is fairly arbitrary).

## Caveats

Although the spatial arrangement procedure is promising, certain problems are peculiar to it. Most importantly, it places fairly strong constraints on subjects'

similarity data. If the constraints imposed by a two-dimensional spatial configuration are not psychologically valid, subjects will not be able to convey their true impressions of similarity. Tversky (1977) has argued that the assumptions imposed by a multidimensional space are violated by empirically obtained similarity data. For example, MDS models predict symmetric similarity. The similarity of  $X$  to  $Y$  should be the same as the similarity of  $Y$  to  $X$ , because the distance from  $X$  to  $Y$  [ $D(X,Y)$ ] in a space is equal to the distance from  $Y$  to  $X$  [ $D(Y,X)$ ]. There are, however, violations of symmetric similarity. For example, North Korea is judged to be more similar to China than vice versa (for another type of violation, see Podgorny & Garner, 1979). Other assumptions of traditional MDS models that are violated include the following: minimality [ $D(A,B) \geq D(A,A) = 0$ ], the triangle inequality [ $D(A,B) + D(B,C) \geq D(A,C)$ ], and the nearest neighbor restriction (for two dimensions, no more than five items can have the same nearest neighbor; see Tversky & Hutchinson, 1986). There are ways to salvage MDS models (Krumhansl, 1978; Nosofsky, 1991), but these violations of a simple correspondence between similarity and distance must be taken seriously. In light of the high correlation between similarity ratings and the spatial arrangement data, we can argue that, for the present set of stimuli, the constraints imposed by a spatial approach to similarity are not too strongly violated.

A second potential difficulty with the spatial arrangement technique is the existence of individual differences in the interpretation of the instructions. While a majority of subjects treat item distance as a continuous measure of similarity, others arrange items into fairly discrete "clumps." Additional instructions and examples have proved useful in inducing subjects to produce continuously varying distances.

Finally, the measure developed here may not be appropriate for all uses. It seems likely that there are different kinds of similarity (Medin et al., 1993). Canes are visually similar to pool cues but are conceptually more similar to wheelchairs. The spatial arrangement task appears to tap into a level of similarity that is relatively cognitive as opposed to perceptual. Similarities from the spatial arrangement task and the rating task seem to be influenced by abstract commonalities such as "fancy letters" and "highly unusual font." Similarity data from the *same/different* judgment task seems to be more perceptually based. Thus, the present technique produces data that are likely to be useful for predicting induction (if  $X$  is similar to  $Y$ , properties of  $X$  will be inferred to be true of  $Y$  also), problem solving (if problem  $X$  is similar to problem  $Y$ , strategies used for solving  $X$  may be used for solving  $Y$ ), and memory (if  $X$  is similar to stored memory  $Y$ ,  $X$  may serve as a good cue for retrieving  $Y$ ). The technique may be somewhat less applicable than confusion measures to predictions of perceptual phenomena such as texture segregation (if  $X$  is similar to  $Y$ , it will be hard to say

where the boundary between  $X$  and  $Y$  is) and feature search (if  $X$  is similar to  $Y$ , it will be hard to locate a single  $Y$  example in an array of  $X$ s).

### Applications

The technique of spatial arrangement is well adapted to particular applications in experimental psychology. First, as suggested above, the technique is probably better suited than confusion measures for applications that are directed toward cognitive, as opposed to perceptual, similarity. Second, the technique is ideal for applications that involve creating MDS solutions. Similarities between all pairs of items are generated—exactly the data needed for an MDS solution. In addition, the constraints imposed by MDS solutions are also imposed by the spatial arrangement technique. If the data satisfy the assumptions of spatial arrangement technique (such as symmetric similarity), they will also satisfy the assumptions of MDS.

The technique is also valuable when individual differences in perceived similarity are expected to be extensive. An individual's behavior in a task that is hypothesized to depend on similarity can be modeled by using the individual's own similarity data. For example, a researcher can model a subject's categorization data by examining the subject's similarity data (Nosofsky, 1986). This modeling of individuals is practical because the spatial arrangement measure provides a great deal of similarity data quickly. Even if subjects are in an experiment for only 1 h, both similarity and categorization data can be obtained.

An open question for applications of the technique is, "Should the similarities be transformed to provide useful data?" Specifically, it might be expected that the difference in similarities between objects that are placed 3 cm and 4 cm apart is greater than the difference in similarities between objects that are placed 6 cm and 7 cm apart. We might consider (following Nosofsky, 1991) modeling similarity as

$$S_{ij} = e^{-cd_{ij}},$$

where  $S_{ij}$  is the similarity of items  $i$  and  $j$ ,  $c$  determines the steepness of the exponential curve, and  $d_{ij}$  is the distance (in pixels) between  $i$  and  $j$ . In one use of the spatial arrangement technique to predict categorization accuracy and reaction time, I compared this exponential relation with the simpler relation of  $S_{ij} = 1/d_{ij}$ . The linear relation between similarity and pixel distance predicted categorization performance at least as well as did the exponential relation with values of  $c$  varied between 0.05 and 5.0. However, for other applications, to assume a nonlinear relation between similarity and distance may be desirable.

### Program Availability

The software to run the spatial arrangement procedure can be obtained free of charge. The software, writ-

ten in Think Pascal, was designed for the Macintosh IICI, IISI, IIFX, and Quadra computers running on System 7.0. Black-and-white or color objects to be displayed are stored as PICT resources. To receive a copy of the software, send a Macintosh diskette to the author at the Psychology Department, Indiana University, Bloomington, IN 47405. To receive a copy of the software over e-mail, send your request to the author at [rgoldsto@ucs.indiana.edu](mailto:rgoldsto@ucs.indiana.edu).

#### REFERENCES

- GOLDSTONE, R. L. (1994). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 3-28.
- GOLDSTONE, R. L., & MEDIN, D. L. (1994). The time course of comparison. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **20**, 29-50.
- GOLDSTONE, R. L., MEDIN, D. L., & GENTNER, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology*, **23**, 222-264.
- JAMES, W. (1950). *The principles of psychology* (Vol. 1). New York: Dover. (Original work published 1890)
- KRUMHANSL, C. L. (1978). Concerning the applicability of geometric models to similarity data: The interrelationship between similarity and spatial density. *Psychological Review*, **85**, 450-463.
- MEDIN, D. L., GOLDSTONE, R. L., & GENTNER, D. (1993). Respects for similarity. *Psychological Review*, **100**, 254-278.
- NOSOFSKY, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, **115**, 39-57.
- NOSOFSKY, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, **23**, 94-140.
- PODGORNY, P., & GARNER, W. R. (1979). Reaction time as a measure of inter- and intraobject visual similarity: Letters of the alphabet. *Perception & Psychophysics*, **26**, 37-52.
- ROSENBERG, S., & KIM, M. P. (1975). The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, **10**, 489-502.
- TORGERSON, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika*, **30**, 379-393.
- TOWNSEND, J. T. (1971). Theoretical analysis of an alphabetic confusion matrix. *Perception & Psychophysics*, **9**, 40-50.
- TVERSKY, A. (1977). Features of similarity. *Psychological Review*, **84**, 327-352.
- TVERSKY, A., & HUTCHINSON, J. W. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, **93**, 3-22.
- V & M TYPOGRAPHICAL, INC. (1974). *The type specimen book*. New York: Van Nostrand Reinhold.