Quality & Quantity **27**: 201–209, 1993.

Note

Multidimensional scaling reliability in similarity judgments about environmental sentences

ANA-DELIA CORREA,[1] JOSÉ DÍAZ,[2] ERNESTO SUÁREZ,[3] & BERNARDO HERNÁNDEZ[4]

[1]*Associate Professor in the Department of "Didáctica e Investigación Educativa". Faculty of Philosophy and Sciences of Education. University of La Laguna. Canary Islands, Spain*; [2]*Associate Professor in the Department of "Psicología Cognitiva, Social y Organizacional". Faculty of Psychology. University of La Laguna*; [3]*Associate Professor in the Department of "Psicología Cognitiva, Social y Organizacional". Faculty of Psychology, University of La Laguna*; [4]*Full Professor in the Department of "Psicología Cognitiva, Social y Organizacional". Faculty of Psychology, University of La Laguna.*

Abstract. The aim of this research was to analyze the stability of Multidimensional Scaling (MDS). The investigations on this topic include studies related to the assumptions of linearity and monotonicity and to the validity and reliability of MDS procedures. Centered in this last topic, a test of the reliability of MDS procedure was carried out. We employed a set of complex stimuli: 21 sentences related to environmental field. 40 subjects made similarity judgements about both pairs "sentence A – sentence B" and "sentence B – sentence A", so obtaining one square matrix per subject. Each of these matrices was broken down into two triangular matrices that were scaled separately by INDSCAL. The coordinates of stimuli in both MDS solutions were correlated. The results show a significant correlation between the two solutions.

Key words: Multidimensional scaling, reliability, similarity judgements, sentences comparation.

The generic denomination of multidimensional scaling (MDS) includes a variety of procedures of multivariate analysis set out to obtain the underlying structure of a matrix of empiric data (the objects can be people, brands, places, etc.) and to represent this structure geometrically, in two or more dimensions using linear or non-linear relations.

The MDS is habitually classified in two basic forms: metric and non-metric, although the progressive development of the technique permits to obtain metric representations from ordinal data. This possibility is especially interesting for Social Sciences, given the frequency with which variables of an ordinal level of measurement are analyzed: preferential judgements, agreement, similarity, etc.

The aim of non-metric procedures is technically to find out a set of points where the dissimilarity data are an (ordered) monotonic function of the

distances in a space of minimal dimension. An *r*-dimensional solution would be a configuration of *n* points (stimuli, in the broadest sense) in a space of *r*-dimensions that fits this criterion for a set of data. The "substantive" interpretation of the dimensions found would be based on the stimuli position in this *r*-dimensional space.

The expansion and development of the various techniques of MDS have allowed its application to an extensive range of formats and types of data. Thus, it is possible to work with square or non-square matrices, lower-half or upper-half matrices, full matrices, etc. Depending on the type of data, it is possible to use preferential data, different types of arrangement, similarities, usefulness, etc. This flexibility justifies the growing popularity of MDS and its increasing use in a variety of sciences, subjects, and research problems. A recent work by Green *et al.* (1989) describes different MDS techniques and their uses in research. Also, they offer a wide choice of references and illustrate practically the application of several different scaling programs.

From methodological point of view, there are several studies analyzing MDS robustness. These ones could be classified into three groups. First group of studies analyzes multidimensional solutions stability against violation of the monotonicity assumptions (non-metric MDS) and linearity assumption (metric MDS), and under simulated desetructuring conditions (Shepard, 1969; Green & Rao, 1970; Spence & Ogilvie, 1973; Cohen & Jones, 1974; Green, 1975a). Conclusions of these studies indicate that MDS solutions act on fairly stable way whenever scaling is done into correct dimensionality, and the number of stimuli is adequate in relation to dimensionality. Second group of studies is devoted to analyze the differences that are produced by means of different algorithms of scaling (Green, 1975b), the order of stimuli presentation (Jain & Pinson, 1976), the number of stimuli (Dong, 1983) and outliers (Spence & Lewandowski, 1989). Given the heterogeneity of this second block of studies and its relative scarceness, it is difficult to come to a definitive conclusion regarding the resistance of MDS to the distorting factors analyzed. In spite of this fact, it seems reasonable to admit that MDS solutions are relatively independent of the characteristics of the task and the procedure of analysis used. Third group includes studies about reliability and validity of MDS. Our research is focussed onto last group.

The most of reliability studies follow a test-retest procedure. In one of the earliest studies, a correlation average of 0.71 in ranks of similarity was found (Weksel & Ware, 1967). More recently, Moore & Lehmann (1982) show individual correlations that range between 0.10 and 0.75. Day *et al.* (1976) analyzed the reliability of similarity judgements, making subjects repeat only a small part of the judgements, obtaining high reliability coefficients. How-

ever, Deutscher (1982) reanalyzed part of the same data and did not achieve such optimistic results. Humphreys (1982) analyzed the influence of different processes of obtaining data in test-retest reliability, which does not seem to be affected by this variation.

In this sense, there is an interesting study of Summers & MacKay (1976), among other things because it gave rise to a heated debate in the form of critical commentaries by Wilkes & Wilcox (1977) and the response to these (Summers & MacKay, 1977). The results obtained by Summers & MacKay (1976), focussed on test-retest reliability and convergent validity, are discouraging with respect to the reliability and validity of similarity judgements as a measure of individual perceptions. However, they recognize how inappropiate it would be to generalize such conclusion to other types of stimuli, other instruments for data collection, etc., and also the necessity to develop alternative methods for the analysis of validity.

More recently, Malhotra in his works (1982, 1987) distinguishes the concepts of test-retest reliability and structural reliability, the latter referring to the stability of MDS solutions when objects or scales are embedded in a wider context of similar stimuli. This is called structural perturbation which may be more or less intense.

Moreover, Malhotra analyzes discriminant and convergent validity applying a multitrait-multimethod analysis. His general conclusion is positive as reasonable indices of discriminant and convergent validity, and adequate resistance of MDS to structural perturbation, are obtained. Malhotra also insists on the need for caution when making generalizations about MDS results obtained in a stimulus field to other fields.

The global conclusion that can be made from this subject, perhaps at this time, is not very clear given the divergencies among results. We think more research is necessary using different stimuli, forms of data collection, subjects and trials of alternative formulas to evaluate reliability and validity. This is the aim of this paper.

Our work involves exploring the reliability or stability of an MDS solution, with two basic variations with respect to existing studies. The first of these differences is due to the type of stimuli used in the similarity judgements. Normally in MDS analysis, stimuli such as commercial brands (cars, soft drinks . . .), places (towns, neighbourhoods . . .), people (politicians, actors . . .) and others of a similar character are used. This already represented greater complexity with respect to the stimuli analyzed at first (colours, shapes, distances). We have used as stimuli, relatively short sentences whose content reflects different ideas regarding the environment, its exploitation, conservation, degradation, etc. The possibility of utilising data of this type

Table I. Example of presentation of statements in the questionnaire

SENTENCE TO COMPARE
"Simple and direct solutions to environmental problems do not produce the necessary results"

Comparison Scale

1	2	3	4	5	6	7
Not at all similar	Very slightly similar	Slightly similar	Somewhat similar	Quite similar	Very similar	The same

1. "To solve environmental problems it is necessary to change our style of living" . . .
2. "The creation of green areas goes against growth" . . .
3. "It is very risky to make our future depend on technological success" . . .
4. "The satisfaction and happiness of citizens will only be possible if we progress economically" . . .
5. "The consumer society is inconsistent with a respect for nature" . . .

. . . An so on up to 21 sentences. . . .

in MDS analysis is especially interesting for Social Sciences, as frequently the interest variables are not constructs that can be represented by just one word.

The second difference is determined by the procedure that we have used to evaluate reliability, as we have adopted a system based on the comparison of repeated pairs. The subjects, for each pair of sentences, should evaluate their similarity twice in the same session: first comparing A with B, and later comparing B with A. In this way, two sets of equivalent judgements are obtained, at the same time avoiding some of the inconveniences associated with test-retest, such as experimental mortality and the effects of uncontrolled variables. It is hoped to find a positive and significant correlation between the solutions generated by MDS starting from this two sets of similarity judgements.

Method

Instruments

The instrument for data collection was a questionnaire in which the degree of similarity of a set of sentences was evaluated. To elaborate it, a set of sentences relative to the environment were selected from bibliographic material, press, television and brainstorming sessions. A total of 101 sentences of similar extension and structure were established. Of these 101 sentences, 21 were randomly selected for the questionnaire. In it, each one

of the sentences appears compared to itself and to the other 20 sentences. The degree of similarity between each pair of sentences is evaluated according to a similaritiy scale that ranges from 1 (not at all similar) to 7 (the same). See in the Table I, as an example, the presentation of one of the sentences.

In asking for the comparison of all the sentences with all the others, a square matrix of data 21 × 21 for each subject was obtained: in the diagonal are the comparisons of each sentence with itself (21), in the upper half are the comparisons of each sentence with the rest (210) and in the lower half are the same comparisons but reversing the order of the sentences in each pair (210). Thus, the total of similarity judgements per participant was 441.

Subjects

The questionnaire was completed by 71 participants selected according to a quota sampling that distributed them in a balanced way in function of sex, and of three age groups: less than 25, between 25 and 50, and more than 50 years old.

Procedure

The completion of the questionnaire was carried out in individual sessions or small groups, in agreement with the availability of the participants.

The instructions expressly indicated that the task consisted of comparing a set of pairs of sentences and evaluating their degree of similarity. It was emphasized that the similarity did not refer to the grammatical structure. Instead, it was stressed that the degree of similarity of the ideas expressed in each of the sentences would be evaluated. In the same way, it was emphasized that the degree of personal agreement with the opinions expressed was not asked for. The total time that participants took to complete the questionnaire was approximately 60 to 75 minutes.

With this procedure, for each subject of the sample a square matrix was obtained, where each pair of sentences is compared twice: [AB, BA, AC, CA; AD, DA; etc.]. The judgements of the diagonal [A-A, B-B, etc.] were rejected for the analysis, being verified only if participants gave them the maximum point on the similarity scale. Each individual square matrix was later broken down into two triangular matrices, thus obtaining two subsets of individual matrices: [AB; AC; AD; etc] and [BA; CA, DA; etc]. Each subset was the object of an independent MDS analysis by means of INDSCAL. Finally, in order to evaluate the reliability of MDS configurations, we correlated the coordinates of stimuli in the multidimensional space found in each subset.

Results

Data analysis was divided into three parts: analysis of direct scoring, scaling, and reliability of MDS solution.

Analysis of direct scoring

Before the scaling, some analysis of data was carried out to test, firstly, that the order of presentation of the sentences did not influence the evaluation of similarity; secondly, that the correlation between symmetric pairs was greater than between pairs with different sentences, which would indicate that the subjects' judgements had not been indiscriminate.

The first was tested by an one-sample t-test for each of two pairs with equal sentences. Out of a total of 210 contrasts, 76.6% did not show significant differences. This percentage does not fit a Poisson distribution with an average of 0.01. Despite this distance from the expected distribution, in the most of comparisons the order of presentation of the stimuli did not affect the evaluation of similarity which the subjects made.

To test if the correlation between pairs with same sentences was significantly greater than the correlation between pairs with different sentences, the product-moment correlations between all the pairs of sentences were calculated. The average correlation for equal pairs was 0.41, and for the different pairs, 0.20. A two-sample t-test showed that the average correlation between equal pairs was significantly greater than that of the different ($p < 0.001$).

Scaling

For this analysis, the following subjects were eliminated: those who had left some answers blank, those which did not contain the maximum value (7) in the diagonal of the matrix, and/or those which showed a systematic tendency to repeat the same value of the similarity scale. Thus 11 subjects were eliminated. Of the 60 remaining, 20 subjects were eliminated at random, so the scaling was carried out with 40 subjects. This was done due to a limitation of the scaling program, according to which, $Ne^2 < 18000$ (N: number of subjects; e = number of stimuli) (MDS, 1981).

Two independent scalings for each subset of data were carried out by means of INDSCAL. Each subset was respectively formed by the upper and lower half of the total square matrix obtained for each subject.

In both cases a solution of two dimensions was selected. The percentage of variance of the tridimensional solution did not increase substantially in

Table II. Correlation of the stimulus coordinates in each dimension of the two triangular matrices of data

	Dimension 1 upper half	Dimension 2 upper half	Dimension 1 lower half	Dimension 2 lower half
Dimension 1	1.000			
upper half	*****			
Dimension 2	0.40	1.000		
upper half	$p = 0.036$	*****		
Dimension 1	0.95	0.44	1.000	
lower half	$p = 0.000$	$p = 0.021$	*****	
Dimension 2	0.42	0.84	0.32	1.000
lower half	$p = 0.027$	$p = 0.000$	$p = 0.075$	*****

comparison to the bidimensional solution. Moreover, we did not obtain a higher degree of independence between isolated dimensions when three dimensions were handled. The percentage of variance explained in the bidimensional solution was of 41.3 for the upper half and 32.5 for the lower half. In the upper half the percentage of variance of the first dimension was 78.5 and of the second dimension 21.5. In the lower half, the first and second dimensions explained a percentage of 69.7 and 30.3 respectively.

Reliability of the MDS solution

To evaluate the reliability of the MDS solution, the scores obtained for the 21 sentences in the two-dimensional space corresponding to the upper half of the matrix were correlated with those obtained in the lower half. That is to say, normalized stimulus coordinates for each dimension obtained from the MDS analysis of half of the questionnaire were correlated with the same coordinates obtained in the scaling of the other half. The results of the product-moment correlations are shown in Table II.

The position of the stimuli in dimension "1" of the lower half has a high and significant correlation with their position in the same dimension of the other half ($r = 0.95$, $p = 0.000$). Regarding dimension "2", the positions in both halves correlate slightly less ($r = 0.84$, $p = 0.000$). In other words, the results of the two scalings show a high correlation between equivalent dimensions along both MDS solutions.

Discussion

The most relevant conclusions that we can make from this work make reference to: on one hand, the stability of similarity judgements when com-

plex sentences are used, and on the other hand, the reliability of the multidimensional solutions that are obtained from these judgements. The fact that the correlation between similarity estimations realized in reverse order were highly positive, shows that the subjects maintain similar criteria of comparison throughout the test. Likewise, we have observed that the correlations between dimensions obtained from a half of the matrix and its homologous dimensions, obtained from the other half, are not only significant but also clearly superior to the correlations between different dimensions taken from the same half. In this sense, we can conclude that we have obtained the same dimensions starting from two triangular matrices of similarity judgements. These results confirm the stability of MDS analysis.

Likewise, the correlation between the two dimensions "1" has been higher than the correlation between the two dimensions "2". This is probably owed to the own nature of the scaling procedures, since the strength of the dimensions, understood as percentage of variance explained, decreases progressively. It would be necessary to go deeper into this aspect using scalings in which more dimensions have been obtained, with the aim of testing if this is a stable tendency.

It is observed that similarity judgements are more stable when they are analyzed in the form of "scores generated by MDS" than when they are analyzed as direct scores.

In spite of a predominant tendency towards symmetry in similarity judgements, some significant differences between A-B and B-A comparisons were found. It is necessary to explore the possibility that the stimuli in which those differences were found, present some differential features (e.g.: length or complexity). In any case, it would not alter the general pattern of correlations, nor, consequently, the configuration and interpretation of the MDS solutions.

It would be neccesary to make more researchings to analyze the sensitivity of MDS to stimulus types, to instruments for data collection and to the procedures employed for analysis. In a parallel way, it would be necessary to make general revisions of the results reached in dispersed studies. These revisions, which could adopt the form of meta-analysis, would analyze the reasons for the discrepancies of the results obtained by different authors, and so reach more global conclusions about the matter.

Acknowledgments

This work was partially supported by Project number 105.31.07.89 of the "Dirección General de Universidades e Investigación del Gobierno Autónomo de Canarias".

References

Cohen, H. S. & Jones, L. E. (1974). The effects of random error and subsampling of dimensions on recovery of configurations by nonmetric multidimensional scaling, *Psychometrika* 39: 69–90.

Day, G. S., Deutscher, T. & Ryans, A. B. (1976). Data quality, level of aggregation, and nonmetric multidimensional scaling solutions, *Journal of Marketing Research* 13: 92–97.

Deutscher, T. (1982). Issues in data collection and reliability in marketing multidimensional scaling studies. Implications for large stimulus sets, In R. G. Golledge & J. N. Rayner (eds.), *Proximity and Preference: Problems in the Multidimensional Analysis of Large Date Sets*, Minneapolis: University of Minnesota Press.

Dong, H. (1983). Method of complete triads: An investigation of unreliability in multidimensional perception of nations, *Multivariate Behavioral Research* 18: 85–96.

Green, P. E. (1975a). On the robustness of multidimensional scaling techniques, *Journal of Marketing Research* 12: 73–81.

Green, P. E. (1975b). Marketing applications of MDS: Asessment and outlook, *Journal of Marketing* 39: 24–31.

Green, P. E., Carmone, F. J. & Smith, S. M. (1989). *Multidimensional Scaling. Concepts and Applications*, Massachusetts: Allyn & Bacon.

Green, P. E. & Rao, V. R. (1970). Ratings scales and information recovery. How many scales and response categories to use?, *Journal of Marketing* 34: 33–39.

Humphreys, M. A. (1982). Data collection effects on nonmetric multidimensional scaling solutions, *Educational and Psychological Measurement* 42(4): 1005–1022.

Jain, A. K. & Pinson, C. (1976). The effect of order of presentation of similarity judgments on multidimensional scaling results: An empirical examination, *Journal of Marketing Research* 13: 435–439.

Malhotra, N. K. (1982). Structural reliability and stability of nonmetric conjoint analysis, *Journal of Marketing Research* 19: 199–207.

Malhotra, N. K. (1987). Validity and structural reliability of multidimensional scaling, *Journal of Marketing Research* 24: 164–173.

MDS (1981). *MDS Package User Manual*, University of Edinburgh: Program Library Unit.

Moore, W. L. & Lehmann, D. R. (1982). Effects of usage and name on perceptions of new products, *Marketing Science* 1: 351–370.

Shepard, R. N. (1969). Some principles and prospects for the spatial representations of behavioral science data. Advanced Research Seminar on Scaling and Measurement, Newport Beach, California.

Spence, I. & Lewandowsky, S. (1989). Robust multidimensional scaling, *Psichometrika* 54(3): 501–513.

Spence, I. & Ogilvie, J. C. (1973). A table of expected stress values for random rankings in nonmetrics multidimensional scaling, *Multivariate Behavioral Research* 8: 511–518.

Summers, J. O. & MacKay, D. B. (1976). On the validity and reliability of direct similarity judgments. *Journal of Marketing Research* 13: 289–295.

Summers, J. O. & MacKay, D. B. (1977). On establishing convergent validity: A reply to Wilkes and Wilcox, *Journal of Marketing Research* 14: 263–265.

Weksel, W. & Ware, E. (1967). The reliability and consistency of complex personality judgments, *Multivariate Behavioral Research* 2: 537–541.

Wilkes, R. E. & Wilcox, J. B. (1977). On the validity and reliability of direct similarity judgments: A comment, *Journal of Marketing Research* 14: 261–262.