Neuropsychology
2003, Vol. 17, No. 2, 321–322

0894-4105/03/$12.00 DOI: 10.1037/0894-4105.17.2.321

Question Hard, Answer Simply: A Comment on Storms et al. (2003)

Yoshio Takane
McGill University

It seems clear that, for whatever reasons, the dementia of the Alzheimer type patient group (as well as other patient groups) exhibits behavior that is different from the normal control group. G. Storms, T. Dirikx, J. Saerens, S. Verstraeten, and P. P. De Deyn (2003) rightfully argue that the observed behavior (similarity judgments) does not tell us the source (cause) of the differences between the 2 groups. Rather, the focus of the study should be placed more on finding the ways the 2 groups are different. They also point out various methodological problems in some of the previous attempts to characterize the nature of the differences. Further methodological issues in G. Storms et al.'s study are examined.

It seems obvious that the kinds of similarity data collected by Chan and colleagues (e.g., Chan, Butters, Paulsen, et al., 1993; Chan, Butters, Salmon, & McGuire, 1993) do not have enough information to determine what kind of deficit patient groups are undergoing. This should be clear from the Warrington and Shallice (1979) criteria that state a set of necessary conditions for distinguishing between access disorders and storage deficits. It is the problem of face validity, and this I can readily say even without going through the series of elaborate statistical analyses done by Storms, Dirikx, Saerens, Verstraeten, and De Deyn (2003). No data analysis methods can overcome the lack of relevant information in the data.

It also seems clear, however, that the patient groups exhibit different behavior from that of normal participants. It is interesting to find out in what ways similarity data elicited by the patient groups are different from those obtained from the normal control group. If we understand that Chan, Butters, Paulsen, et al.'s (1993) and others' efforts are primarily aimed at finding the nature of the differences between the groups, their attempts are still worthwhile despite the shortcomings in their methodology as pointed out by Storms et al.

The clear and important message I get from Storms et al's (2003) article is, the simpler, the better. Simpler methods are better suited, if they meet genuine data analytic objectives. Overly sophisticated methods, on the other hand, may unnecessarily complicate the situation. Assessment of reliability in similarity judgments is the case in point. Storms et al. simply take the test–retest reliability between the original similarity data, whereas Chan, Butters, Salmon, and McGuire (1993) first apply individual differences scaling (INDSCAL) to the original data, and then correlate the weights derived from INDSCAL over two occasions. The latter can lead to overly optimistic estimates of reliability, as INDSCAL analysis screens out certain variabilities in the original data that are not consistent with the assumptions of the model it fits.

Multidimensional scaling (MDS) is a collection of powerful data analytic techniques for representing similarity data. However, it typically requires sets of stringent assumptions on the data. There are many pitfalls as well, only some of which are described by Storms et al. (2003). The problem of convergence to nonglobal minima, for example, is something no one cares to talk about these days, but the problem persists as it did 30 years ago (Arabie, 1973). Although by no means do I intend to scare away potential users of MDS in semantic network research, let me reiterate that special care must be taken, particularly when dealing with special groups of participants. Storms et al.'s article sends a warning signal against certain uses of MDS.

Having said the above, I now turn to a few specific remarks on some of the data analyses conducted by Storms et al. (2003).

1. Storms et al. (2003) initially used alternating least squares scaling (ALSCAL) to obtain benchmark distributions of stress values under random rankings. They then compared the stress values obtained from empirical studies done by Chan, Butters, Paulsen, et al. (1993) and others to conclude that the stress values obtained from the patient group were not significantly different from those obtained from the random ranking studies. But why ALSCAL? This is a bit peculiar, inasmuch as ALSCAL does not minimize stress, and consequently obtains stress values larger than those obtained by the methods that explicitly minimize stress.

2. Storms et al. (2003) generated random triadic comparison data and analyzed them along with the actual normal participant data by INDSCAL. The results were used to show that Chan, Butters, Paulsen, et al.'s (1993) finding that the DAT patient and control groups were well discriminated on the basis of the weights obtained by INDSCAL was a mere artifact of the analysis. More specifically, they showed that an equally good discrimination could be made between the normal participant weights and those obtained from the random triadic comparison data. However, the random tri-

This work was supported by Grant A6394 from the Natural Sciences and Engineering Council of Canada to Yoshio Takane.

Correspondence concerning this article should be addressed to Yoshio Takane, Department of Psychology, McGill University, Montreal, Quebec, Canada H3A 1B1. E-mail: takane@takane2.psych.mcgill.ca

adic comparison procedure necessarily creates homogeneous data sets with a lot of similarity values concentrated near averages, which are characteristically different from the normal participant data. It is no surprise then that Storms et al. could get such a good discrimination between the normal and random patient groups. The key question is whether the actual patient data resemble the random triadic data in their degrees of homogeneity.

3. Storms et al. (2003) repeated the random ranking studies using the data sets they collected themselves. To show that the stress values obtained from MDS analyses of aggregated data were not significantly different from those obtained from random ranking data, they used Spence and Ogilvie's (1973) criterion. But why? It seems that this procedure is based on the assumption that the stress values follow an approximate normal distribution around their means. This is not likely to be true, inasmuch as they are bounded. Particularly, they are bounded from below by 0, which is rather crucial because the range of stress values concerned are close to the lower bound, and the distributions of the stress values in this range are most likely to be skewed. It makes much more sense (and is simpler as well) to directly calculate the p values corresponding to the observed stress values.

4. Storms et al. (2003) compared INDSCAL solutions obtained by Chan, Butters, Salmon, and McGuire (1993) and those obtained by themselves for the control groups and the patient groups. They found that the mean agreement between stimulus configurations obtained from the control groups was substantially higher than that between the patient groups. In the process of this analysis, they transformed one of the two configurations to be compared with a maximum agreement by scaling, rotation, and reflection. Stimulus configurations derived from INDSCAL are, however, uniquely determined up to reflection of the coordinate axes. They should not have allowed any transformations other than reflection. It is also not clear if the degrees of agreement between the patient configurations is within the range of a chance level. Somewhat unfortunately, no random ranking studies were conducted for INDSCAL analysis.

Storms et al. (2003) portray the INDSCAL model as a very restrictive one requiring the existence of common dimensions across individuals and allowing only differential weightings of those dimensions by different individuals. However, the fact of the matter is contrary. In fact, it is as general as separate analyses of individual data by the simple (unweighted) Euclidean model. Suppose there are two stimulus configurations having no commonality between them. That is, $\mathbf{X}'_1\mathbf{X}_2 = 0$. The two configurations can, however, be put in a single weighted Euclidean model by joining the two configurations together by $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]$, and by defining the weight matrices of the two individuals by $\mathbf{W}_1 = \text{diag}(\mathbf{I}, 0)$ and $\mathbf{W}_2 = \text{diag}(0, \mathbf{I})$.

5. Storms et al. (2003) applied separate MDS analyses of individual data by the simple Euclidean model in view of the fact that there was little inter-individual consistency in DAT patients data. I would argue, however, that the procedure used to derive similarity data from triadic comparison data invariably creates many tied observations at an individual level. There are $n(n-1)$ pairs of stimuli, where n indicates the number of stimuli, and there are only $n-1$ possible distinct similarity values. Random ranking studies should take into account the numbers of tied observations in the observed data, as they may potentially have a significant impact on the distributions of the stress values. Preferably, the random ranking data contain equal amounts of tied observations as the observed similarity data. One plausible way of ensuring this is to randomly permute the observed data to obtain the random ranking data. Existence of so many tied observations may account for part of the reason why high degrees of misfit were observed in some of the normal participants' data.

6. Assessment of the reliability is very essential in any kind of data analysis, and I am glad that Storms et al. (2003) did it the right way. As has been mentioned already, it was good that they correlated original similarity data over two occasions. They could have done better, however, if they had collected the test–retest data from the normal control participants as well. They then could have compared the range of test–retest reliability across the two groups. They could have done even better, if they had collected the data not only twice but three times. They then could have compared the range of reliability across different intervals of time between measurements.

It is striking to find such wide ranges of individual differences (as well as group differences) in reliability and in the degree of representability of individual similarity data by the Euclidean distance model. A univariate analysis of variance may help in tearing apart various effects attributable to different sources. Another notable observation is that there seem to be some degrees of reliable components in the patient data, no matter how small they may be. It is tempting then to find out what they are.

References

Arabie, P. (1973). Concerning Monte Carlo evaluations of nonmetric multidimensional scaling algorithms. *Psychometrika, 38,* 607–608.

Chan, A. S., Butters, N., Paulsen, J. S., Salmon, D. P., Swenson, M. R., & Maloney, L. T. (1993). An assessment of the semantic network in patients with Alzheimer's disease. *Journal of Cognitive Neuroscience, 5,* 254–261.

Chan, A. S., Butters, N., Salmon, D. P., & McGuire, K. A. (1993). Dimensionality and clustering in the semantic network of patients with Alzheimer's disease. *Psychology and Aging, 8,* 411–419.

Spence, I., & Ogilvie, J. C. (1973). A table of expected stress values from random rankings in nonmetric multidimensional scaling. *Multivariate Behavioral Research, 8,* 511–517.

Storms, G., Dirikx, T., Saerens, J., Verstraeten, S., & De Deyn, P. P. (2003). On the use of scaling and clustering in the study of semantic deficits. *Neuropsychology, 17,* 289–301.

Warrington, E. K., & Shallice, T. (1979). Semantic access dyslexia. *Brain, 102,* 43–63.

Received September 16, 2002
Revision received October 20, 2002
Accepted October 20, 2002 ■