Neuropsychology
2003, Vol. 17, No. 2, 318–320


0894-4105/03/$12.00 DOI: 10.1037/0894-4105.17.2.318

# Is There Madness in the Method? A Comment on Storms et al. (2003)

Timothy T. Rogers
Medical Research Council, Cognition and Brain Sciences Unit

G. Storms, T. Dirikx, J. Saerens, S. Verstraeten, and P. De Deyn (2003) offer a critique of multidimensional scaling studies purporting to reveal systematic semantic deficits in Alzheimer's disease and other syndromes which, in contrast to prior claims, demonstrates that patient performance is indiscriminable from random data. The argument is an indictment against the conclusions drawn by previous researchers; but a consideration of the materials used in these studies suggests that the problem may not lie in the method per se, but in poor stimulus design. In all of the similarity judgment tasks reviewed, stimulus items were differentiated solely on the basis of subtle conceptual distinctions of the sort known to be most vulnerable to semantic impairment. Thus, one would not expect systematic responding from semantically impaired patients to begin with. An alternative design is suggested.

Storms, Dirikx, Saerens, Verstaeten, and De Deyn (2003) have put forward a thorough indictment of several recent multidimensional scaling (MDS) studies designed to investigate semantic memory impairment, primarily in Alzheimer's disease (AD) but also in schizophrenia and Williams syndrome. The arguments marshaled by the authors persuade me that the data under question are effectively uninterpretable and, thus, do not help to adjudicate whether the patient groups under consideration have semantic memory impairments. I therefore focus on two questions. First, are the data uninterpretable because the MDS methods scrutinized by the authors are generally unsuited to analyzing semantic impairments in these patient groups, or do the difficulties brought to light in the article arise instead from the details of the particular experiments reviewed by the authors? Second, how might one assess the validity of the method, in order to better understand why the data it has yielded are so uninformative?

To begin, it will be useful to consider why, in my view, the case against the MDS data is persuasive. The authors target MDS studies in which proximity matrices have been derived from sorting, fluency, and similarity judgment tasks; however the latter two thirds of the critique focuses on the similarity judgment task, which has heretofore been understood to produce the most reliable and informative data for a variety of reasons (see, e.g., Chan et al., 1995, in addition to Storms et al., 2003). I will therefore restrict my own comments to the data from this task.

The similarity judgment requires participants to determine, for all possible triplets in a corpus of words, which two of three concepts are most similar in meaning. From these judgments, pairwise proximity matrices are constructed, which can be decomposed in various ways to assess the factors underlying the detection of similarity (for the group as a whole or for an individual relative to a normative frame of reference). Storms et al. (2003) present a thorough set of analyses designed to show that, under any method of analysis or interpretation, MDS solutions are less suited to explaining patient data than normal control participant data. There are three broad planks to the authors' argument, two of which I do not find especially convincing and the third of which I think is critical.

First, the authors showed that MDS group data fits are much poorer for patient groups than for control participants, both in published studies and in their own data. The exercise is interesting, but this is hardly front-page news. Regardless of syndrome, patient data is inevitably more variable than control data. Even if individuals in the patient groups had qualitatively and quantitatively similar impairments, one would expect somewhat more random behavior and, hence, worse-fitting MDS solutions, in patients compared with control participants. That such an effect is observed in the present case does not in itself provide evidence against the utility of MDS techniques for interpreting patient data.

Second, the authors showed that an individual patient is not as likely as a normal control participant to give the same pattern of responses on two different testing occasions. The authors suggested that this finding is incompatible with storage-based accounts of patient deficits, which they argued should predict consistent patterns of responding over time. Considering the demands of a forced-choice task, however, reduced reliability in the patient data is not surprising under any theory of the deficit. If, on a given trial, none of the three words presented to the patient is meaningful (as a consequence of semantic impairment), the patient must guess at random. Any such trials will introduce noise into the pattern of performance witnessed on different testing occasions and will thus reduce the individual's test–retest reliability. The data reported by the authors show that, for all but one AD patient, an individual's score on one test session correlates no better with his own score on an earlier session than with the group average. From this result, it is possible to conclude only that intraindividual behavior in the task is no more systematic than interindividual behavior.

Correspondence concerning this article should be addressed to Timothy T. Rogers, Medical Research Council, Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 2EF, United Kingdom. E-mail: tim.rogers@mrc-cbu.cam.ac.uk

There may be certain trials for which all patients make reliable judgments and certain trials for which all patients guess at random. In such a case, the test–retest reliability for a single patient would not be expected to exceed the inter-individual reliability of the measure.

For me, the persuasiveness of the article rests on the third plank of the authors' argument: the demonstration that the observed patient data is indistinguishable from random data. The demonstration is crucial because it indicates that there is no statistical basis for concluding that the patients, taken individually or as a group, are behaving in any systematic fashion whatsoever. The argument seriously undermines claims made by Chan et al. (1995) about particular patterns of spared and impaired knowledge observed in these syndromes.

## Is It the Method or the Materials?

Why is patient behavior in these studies indistinguishable from random performance? The authors correctly point out that there could be many reasons, including attention deficits, executive disorders, and semantic access disorders. However, the literature is not mute with respect to determining which aspects of semantic knowledge are most likely to be robust to so-called *storage deficits* and which most vulnerable. As Storms et al. (2003) noted, a broad range of studies have demonstrated that knowledge about broad conceptual distinctions (e.g., animals vs. artifacts) is considerably more robust to semantic storage deficits than is knowledge about more fine-grained distinctions (e.g., goat vs. pig, boat vs. plane; see Done & Gale, 1997; Hodges, Graham, & Patterson, 1995; Hodges, Salmon, & Butters, 1991; Warrington, 1975). From these observations, one would not expect patients with a degraded semantic store to perform other than randomly in the triad tasks reviewed in Storms et al.'s (2003) critique, because all of these studies required participants to make similarity judgments among triads of closely related items: four-legged land animals differentiated by subtle dimensions such as predation and domesticity (one study also tested a set of tools, another narrow semantic category). Given that patients with semantic impairments can have considerable difficulty discriminating ducks from sheep (e.g., Hodges et al., 1995), it seems unreasonable to expect them to reliably discriminate more closely related items, such as lions, bears, and pigs.

It is possible, then, that the patients studied by Chan et al. (1995) behaved effectively at random in the triad tasks because the stimuli with which they were tested were not differentiated along any conceptual dimension likely to be spared under semantic impairment. Thus, it may be that the MDS methods of interest are perfectly suited to analyzing data from AD and other patient groups, but that previous researchers have held somewhat misguided intuitions about which conceptual distinctions would prove robust to semantic impairment and, consequently, have used stimuli that were not optimal for assessing the structure of degraded semantic representations.

## How to Test the Method

Of course, there are many other reasons why the patients might have performed at random on these tasks. I have two suggestions for determining whether the method itself has any external validity and how it might be used to more fruitfully assess the nature of impaired semantic task performance in AD and other disease groups.

First, the task should use items that are likely to yield predictable results on the basis of what is known about semantic impairment generally. The triad task is limited in the number of concepts that can be assessed in a single session, but even with the standard dozen, it should be possible to use items that vary along two dimensions, one of which is likely to be preserved under semantic impairment, and one compromised. For example, one might use six animals (half water creatures and half air creatures) and six vehicles (half watercraft and half aircraft). The living–nonliving distinction is known to be less vulnerable to semantic impairment than is the water–land distinction (e.g., Hodges et al., 1995). Patients with semantic deficits should thus perform reliably for trials in which items differ on the preserved dimension (e.g., jet, lorry, duck), more randomly when the relevant distinction is degraded (e.g., jet, aeroplane, lorry), and at random when distinctions are very subtle (e.g., jet, aeroplane, blimp). If this prediction were borne out, one would expect MDS analyses to show stress values somewhat higher than those of normal control participants but significantly different from random data.

Second, it would be useful to pretest any such task in patients known to have semantic storage deficits, such as those suffering from semantic dementia (SD), a progressive disorder marked by the profound loss of conceptual knowledge and abilities, coincident with a remarkable sparing of other cognitive faculties. A long tradition of research has documented several aspects of structure in the behavior of such patients tested with a variety of semantic tasks. Most important to the current work, the pattern of relative preservation to superordinate knowledge and distinctions, with the rapid degradation of more specific knowledge and distinctions, has been reliably observed in a very large number of cases and modalities of testing (see Patterson & Hodges, 2000, for a recent review). For example, patients with SD can often produce general, but not specific, names for objects (e.g., animal instead of bird or chicken); are better at sorting both words and pictures into general rather than specific categories; and omit specific but not general properties from their definitions (Hodges et al., 1995). The absence of comparative MDS data from such a patient group is an obvious hole in the current literature—the external validity of the method has yet to be assessed. If the method fails to yield useful data even from patients with well-established semantic storage deficits, it is unlikely to do better in questionable cases. Testing with a group of SD cases could serve both to establish the method's validity and to provide a frame of reference for interpreting the data from other patient groups.

If the method's validity can be firmly established, and predictable nonrandom outcomes are obtained from a pa-

tient group with well-established semantic storage deficits, then it should be possible to draw useful conclusions about the nature of impaired semantic task performance in questionable syndromes such as dementia of the Alzheimer type using these MDS techniques. Specifically, if AD patients continue to behave at random under the proposed conditions, there would indeed be little evidence from MDS that such patients have storage deficits. If AD behavior is found to be nonrandom, but MDS models provide a worse fit for AD than for SD, this would be consistent with a hypothesis of storage deficits with additional complicating factors in AD. If MDS models for AD and SD have comparable fits (and both groups perform nonrandomly), this would provide evidence that deficits in both syndromes arise from storage problems.

## References

Chan, A. S., Butters, N., Salmon, D. P., Johnson, S., Paulsen, J., & Swenson, M. (1995). Comparison of the semantic networks in patients with dementia and amnesia. *Neuropsychology, 9,* 177–186.

Done, D. J., & Gale, T. M. (1997). Attribute verification in dementia of Alzheimer type: Evidence for the preservation of distributed concept knowledge. *Cognitive Neuropsychology, 14,* 547–571.

Hodges, J. R., Graham, K., & Patterson, K. (1995). Charting the progression of semantic dementia: Implications for the organization of semantic memory. *Memory, 3,* 463–495.

Hodges, J. R., Salmon, D. P., & Butters, N. (1991). The nature of the naming deficits in Alzheimer's and Huntington's disease. *Brain, 114,* 1547–1558.

Patterson, K., & Hodges, J. (2000). Semantic dementia: One window on the structure and organization of semantic memory. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology: Vol. 2. Memory and its disorders,* (2nd ed., pp. 313–333). Amsterdam: Elsevier Science.

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. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology, 27,* 635–657.

Received September 16, 2002
Revision received October 17, 2002
Accepted October 17, 2002 ■