
Article:
Prescott, T.J., Newton, L.D., Mir, N.U. et al. (2 more authors) (2006) A new dissimilarity measure for finding semantic structure in category fluency data with implications for understanding memory organization in schizophrenia. Neuropsychology, 20 (6). pp. 685-699. ISSN 0894-4105

https://doi.org/10.1037/0894-4105.20.6.685

Running head: SEMANTIC STRUCTURE IN CATEGORY FLUENCY DATA

A new dissimilarity measure for finding semantic structure in category fluency data with implications for understanding memory organization in schizophrenia

Tony J. Prescott, Lisa D. Newton, Nusrat U. Mir,

Peter W. R. Woodruff, and Randolph W. Parks

University of Sheffield, UK

Abstract

The ordering of words in category fluency lists is indicative of the semantic distance between items in conceptual memory. Several studies have concluded from structural analyses of such data, using cluster analysis or multidimensional scaling, that the semantic memory of patients with schizophrenia is more disorganized than that of controls. Previous studies have based their analyses on a measure of average inter-item dissimilarity devised by Chan et al. (1993). Here we derive a new and improved method of determining dissimilarity and show that when this measure is applied to the fluency lists of patients with schizophrenia the average pattern of organization for the animal category has similar structure to that of controls, but with greater variability between individuals.

Keywords: category fluency, cluster analysis, multidimensional scaling, semantic memory, schizophrenia

A new dissimilarity measure for finding semantic structure in category fluency data with implications for understanding memory organization in schizophrenia

In the category fluency task participants are asked to name as many exemplars as possible of a given semantic category in a short space of time. Data from this task, typically using the category 'animals', have been the subject of many studies—with adults, children, and clinical populations—aimed at elucidating the structure of semantic memory (e.g. Gruenewald & Lockhead, 1980; Henley, 1969; Wixted & Rohrer, 1994), determining how this structure changes during normal development (see Crowe & Prescott, 2003), or becomes disorganized through mental illness (see below).

A pioneering study into semantic memory deficits in clinical groups was performed by Chan and her co-workers (Chan et al., 1993) in patients with Alzheimer's disease. These researchers used the category fluency output of patients, and of age-matched normal controls, to construct representations of semantic memory organization using the statistical methods of *hierarchical cluster analysis* (HCA) and *multi-dimensional scaling* (MDS). A key contribution of that article was the development of an algorithm to derive a measure of the *dissimilarity*, or *semantic distance*, between pairs of concepts based on how frequently the two items appeared together in participant's fluency lists, and the typical distance (i.e. number of intervening items) between the two items in those lists. By calculating this measure for all possible pairs of a given set of category items the average pattern of conceptual organization for this set can be reconstructed, either as a hierarchical tree (using HCA), that links items into clusters and clusters into super-ordinate groups, or as a 'cognitive map' (using MDS) in which relationships between items are indicated by their relative positions in space. The Chan et al. dissimilarity measure, together with these statistical methods for finding structure in group data, were subsequently used to investigate memory

organisation in various clinical populations, and, in particular, in several studies of patients with schizophrenia (Aloia, Gourovitch, Weinberger, & Goldberg, 1996; Moelter et al., 2001; Paulsen et al., 1996; Rossell, Rabe-Hesketh, Shapleske, & David, 1999; Sumiyoshi et al., 2001; Sumiyoshi et al., 2005a, b). All of these studies have argued that conceptual memory organization in patients suffering from schizophrenia is more disorganized than that of controls. For instance, Aloia et al. (1996) claimed that patients with schizophrenia "appear to lack organization and logical associations within their semantic network" (p. 271), while Sumiyoshi et al. (2001) claimed evidence for a "deterioration of the semantic structure […] that is specific to patients with schizophrenia, and is not affected by cultural backgrounds" (p. 196). This reported loss of organization has been variously argued to include the disintegration of super-ordinate categories, the breakdown of strong inter-item associations, the disappearance of organizing dimensions corresponding to concept properties, and the emergence of bizarre or unusual concept groupings.

The goal of the current article is to re-assess the evidence for memory disorganization in schizophrenia as revealed by the category fluency task. Our starting point will be an analysis of the problem of inferring semantic distance from fluency data, leading to the derivation, from first principles, of a new measure of inter-item dissimilarity. On the basis of this analysis we will then review aspects of the methodology used in previous studies and identify two significant limitations affecting most prior research. First, we will show that the prevailing measure of item dissimilarity, proposed by Chan et al. (1993), provides a sub-optimal method for deriving estimates of semantic distance. We will demonstrate that this measure only partially corrects for differences in list length between participants, and is distorted, as a measure of semantic distance, by differences in item frequency. Second, we will argue that by *not* controlling for differences in average list length between groups the data-sets used in some previous studies have over-sampled the fluency output of controls

relative to that of patients. This over-sampling may have resulted in reconstructions of memory organization in which the category structure of patients appears differently organized simply because it is based on less data.

Following our analysis and review of methodological issues, we will present category fluency data from patients with schizophrenia and age-matched controls that are balanced to control for differences in list-length between groups. We will then apply our newly-derived measure of inter-item distance to this data-set and generate structural (cluster and scaling) analyses of semantic memory organization that can be compared across groups and contrasted with analyses computed using the Chan et al. measure. We will show, on the basis of these analyses, and of other statistical investigations of the dissimilarity data, that conceptual memory in schizophrenia, as revealed by the fluency task, is less disorganized than suggested in previous studies.

Patients with schizophrenia reliably generate fewer items in the category fluency task than controls (see Bokat & Goldberg, 2003, for a meta-analysis of relevant studies). However, there is considerable controversy concerning the extent to which this result, and findings of other semantic memory impairments in schizophrenia, arise from a specific problem in memory organization or from a general cognitive impairment leading to difficulties in accessing or searching a lexical store that is more-or-less intact (for recent surveys of this literature see Bokat & Goldberg, 2003; Elvevåg, Weinstock, Akil, Kleinman, & Goldberg, 2001; Giovannetti, Goldstein, Schullery, Barr, & Bilder, 2003; Green, Done, Anthony, McKenna, & Ochocki, 2004; Vinogradov et al., 2003). Structural analyses of data from the category fluency task can contribute to this debate by helping to characterise the ways in which patient fluency data deviate from that of controls over and above these basic differences in fluency output. A re-assessment of the methodology used in such studies could

therefore have significant implications for the broader understanding of semantic memory impairment in schizophrenia.

Derivation of a Novel Measure of Inter-item Distance

Verbal fluency data consist of lists of words generated, in production order, by each participant. The serial nature of speech ensures that, to the extent that fluency data reflect conceptual organization, the high-dimensional structure of semantic memory is compressed to a one-dimensional sequence of words. To recover useful information about the underlying semantic structure many sequences are therefore required, each providing a complementary sample of the underlying representation. Although these multiple samples can be taken from single participants tested on successive occasions, the more usual practice is to acquire fluency lists from many participants and hope that similarities in memory organization will emerge when the data are pooled.

The raw data in the fluency task, from which a measure of semantic distance can be derived, are the set of *inter-item distance pairs*. If we let *a* and *b* be the two items occurring at index positions (using the positive integers 1, 2, 3, …) i_{al} and i_{bl} in a given listing *l* then we can calculate the 'raw' inter-item distance d_{abl} as the absolute difference of these values $d_{abl} = |i_{al} - i_{bl}|$. The distance d_{abl} is zero between an item and itself, one between adjacent items, two for a pair separated by just one intervening item, and so forth. So, for example, in the list [*dog cat rabbit*] the pairs [*dog cat*] and [*cat rabbit*] will be at distance one, and [*dog rabbit*] at distance two. The task of extracting useful semantic distance estimates from such data is complicated, however, by the problem that different participants (or the same individual tested at different times) generate lists of varying length. In an appendix to this article provided on the Supplemental Data Web Page we show that these differences in list length are important, and derive a principled method for combining

measures of inter-item distance across multiple fluency lists that takes due account of them. This normalized distance measure, which is based on a calculation of *cumulative frequency* can be written as follows

$$D(d_{abl},n_l) = (2d_{abl}n_l - d_{abl}^2 - n_l)/n_l(n_l - 1). \quad (1)$$

where n_l denotes the number of items in the list *l*.

Participants in fluency tasks, and particularly those with memory difficulties, may also mention specific items more than once in their fluency lists. Therefore, we also derive (see Supplemental Data Web Page) a modified version of (1) to deal appropriately with lists containing repetitions. Specifically, we select our 'raw' measure of inter-item distance $\hat{d}_{abl} = \min_{\forall a,b \in l} d_{abl}$ to be the smallest of all such distances, define o_{al} and o_{bl} to be the number of occurrences in *l* of items *a* and *b* respectively, and define the weighted distance measure

$$\hat{D}_{abl} = \left[D(\hat{d}_{abl},m_{abl})\right]^{\lambda_{abl}} \quad (2)$$

where $m_{abl} = n_l - (o_{al} + o_{bl} - 2)$ *and* $\lambda_{abl} = \left(o_{al}o_{bl}\right)^{-2/3}$. Hence, for example, for the pair [dog cat] in the list [dog cat rabbit dog], *d*= 1 (minimum of 1 and 2), *m*= 4 – (2 + 1 -2)= 3, and $\lambda = \left(2\right)^{-2/3} = 0.63$. Note that when neither a or b is repeated in a given list Equation 2 reduces to Equation 1, even if the list contains repetitions of other items.

Combining estimates of inter-item distance across participants

The group estimate of the dissimilarity between two items, *a* and *b*, can now be obtained by combining the normalized distances $\hat{D}_{abl}$ calculated from individual fluency lists. Letting T_{Gab} be the total number of *co-occurrences* of items *a* and *b* (i.e. the number of lists

for group *G* that contain both items), we define the *mcf*, or *mean cumulative frequency*, metric such that

$$\text{mcf}(G,a,a) = 0, \quad \text{mcf}(G,a,b) = \frac{1}{T_{Gab}}\left(\sum_{l \in G;a,b \in l} \hat{D}_{abl} \right). \tag{3}$$

In other words, mcf*(G,a,b)*, is a measure of inter-item dissimilarity for group *G* that is zero (maximally similar) between an item and itself, greater than zero for all non-identical items *a* and *b*, and increases (towards a maximum bound of 1.0) with the normalized inter-item distance between *a* and *b* averaged over all group members who named both items at least once in their category list. Finally, we define the set of items A and the dissimilarity lower-half matrix $\mathbf{MCF}_{G,A}$ with elements mcf*(G,a,b)* corresponding to all possible combinations of $a,b(a \neq b) \in A$ found in the fluency lists of group *G*.

Comparison with the Chan et al. Measure of Inter-item Dissimilarity

Next we compare the *mcf* metric with a dissimilarity measure proposed by Chan et al. (1993) and used in all subsequent structural analyses of category fluency data for clinical groups. Using the notation developed above, this prevailing measure, which we will refer to for convenience as the *dis* (for dissimilarity) metric, can be written as

$$\text{dis}(G,a,a) = 0, \quad \text{dis}(G,a,b) = \frac{N_G}{T_{Gab}^2}\left(\sum_{l \in G;a,b \in l} \frac{d_{abl}}{n_l} \right), \tag{4}$$

where N_G is the total number of lists generated by the participant group. Using this metric we can also calculate the dissimilarity half-matrix $\mathbf{DIS}_{G,A}$ defined over the set of items *A* for group *G*.

The *dis* and *mcf* metrics differ both in the way that distances are calculated *within* the list of a single participant, and in how they are combined *across* participants. In an appendix provided via the Supplemental Data Web Page we consider both sources of difference and demonstrate that both introduce a degree of distortion to the *dis* metric. First, the *dis* measure uses division by list length (n_l) to generate a within-list, normalized distance measure; we show that this normalization method significantly under-estimates the correct frequency-based measure (Equation 1) across much of its range. Second, in combining estimates across participants, we can see from Equation 4 that the *dis* metric does not take a simple average of the set of normalized inter-item distances (this would be division by T_{Gab} as in Equation 3) but instead weights this average by N_G/T_{Gab} (giving N_G/T_{Gab}^2 overall). We show that this weighting introduces a confound of the intended measure of inter-item dissimilarity with item *production frequency* such that, all else being equal, pairs composed of high frequency items are accorded smaller inter-item distances than pairs composed of medium or low frequency items. Thus, for example, the pair [*dog elephant*] could end up being rated as more similar, as a result of being named in the animal lists of most participants, than the pair [*cheetah leopard*] which will likely be named by just a few. Finally, it is also worth noting that the *dis* metric does not make any allowance for repeated items within lists, an omission that could have significant consequences if the clinical population of interest shows greater levels of perseveration than controls.

Balancing Sample Size when Comparing Patient and Control Groups

The above analysis indicates that the dissimilarity measure used in all previous structural analyses of category fluency data from clinical groups has several deficiencies that may lead to biased estimates of inter-item dissimilarity. There is therefore a strong case for re-evaluating clinical data from this task using the newly-formulated *mcf* metric to determine

whether the semantic disorganization observed in earlier studies persists when this method of distance estimation is used. Before doing so, however, it is worth noting a second methodological issue, affecting most previous studies, which is the balance (or imbalance) in the amount of data used when making comparisons between patients groups and controls. Specifically, although all studies that have performed structural analyses of fluency data from patients with schizophrenia have shown that patients generate fewer words than controls, this difference in output is often *not* controlled for when calculating average inter-item dissimilarity. In several studies (Aloia et al., 1996; Moelter et al., 2001; Elvevåg and Storms, 2003, Sumiyoshi et al., 2005a) the number of control participants has been equal to or greater than the number of patients, which means that analyses for the latter group will have been based on far fewer item pairs. For instance, Aloia, et al. (1996) tested 28 patients and 32 controls, selecting 11 high-frequency animal names for detailed investigation and reporting that patients with schizophrenia generated significantly fewer of these target items (mean 5.7) than controls (mean 7.0). These initial imbalances in group size and average fluency output are exacerbated by the way in which the number of raw inter-item distances scales with the number of item combinations. Indeed, from the data provided, we can estimate that the control group in the Aloia et al. study generated approximately 80% more distance measures than the patient group.[1] Based on substantially fewer samples, the dissimilarity matrix for the patient group is likely to have contained noisier estimates of inter-item distance than that for the controls. Differences between the cluster trees or scaling analyses based on these matrices might then be due to this imbalance rather than to any more fundamental difference between the groups. There have only been two studies involving patients with schizophrenia (Paulsen et al., 1995; Sumiyoshi et al. 2001) in which the patient group has outnumbered the control group by an amount sufficient to compensate for this difference in verbal fluency output.

The issue of balancing the number of raw inter-item distances across experimental group is particularly important because the clustering and scaling methods employed in previous studies treat all matrix elements identically (i.e. as carrying equivalent amounts of information). At the same time, however, the nature of category fluency data is such that some item pairings—even of, in themselves, high-frequency items—occur at relatively low frequencies. If estimates for these low frequency pairings are based on a very small number of observations (possibly just one), this can contribute considerable noise to the dissimilarity matrix and may generate substantial distortion in the resulting clusters and maps. As a result of these concerns, in the study reported below, care was taken to ensure that the number of raw inter-item distances was balanced between the control and patient groups (rather than balancing the number of participants in each group). We also provide statistics concerning the number of item pairings on which the structural analyses are based,[2] and directly measure the variance in the dissimilarity estimates generated by patient and control groups. We will show that by *partialing out* the effects of group differences in production frequency in this way we can formulate a clearer picture of the differences in semantic organization between patients and controls, as revealed by the category fluency task.

Method

Participants

Forty patients diagnosed with schizophrenia and twenty-eight normal control volunteers were tested for this study. All participants were male and spoke English as their first language. Exlusion criteria included: neurological disorder, epilepsy, hypertension, previous episode of unconsciousness, significant difficulties in hearing, or an alcohol or drug dependence. Demographic variables for the two participant groups, together with appropriate statistical

comparisons and effect sizes, are summarized in Table 1. Participants were aged between 18 and 47 years. Using an alpha level of 0.05, patients did not differ significantly from controls in age but did have fewer years of formal education.

The patient group was recruited from the caseloads of National Health Service psychiatrists in the Sheffield area, and diagnosed by experienced psychiatrists on the basis of medical history and appropriate tests. All patients met the DSM – IV criteria (American Psychiatric Association, 1994) for chronic schizophrenia. All were medicated, 31 (77.5%) on atypical and 9 (22.5%) on typical antipsychotics. Of those treated with atypicals, 17 were receiving olanzapine, 9 clozapine, 3 risperidone, and 2 quetiapine. The chlorpromazine equivalent mean daily dose across all patients was 474.0 mg (s.d. = 339.3). The age of first diagnosis ranged from 14-45, with average age of onset 23.45 years (s.d. 6.92). 32.5% were diagnosed before age 20, and 97.5% before the age of 45. The mean time since diagnosed was 9.33 years (s.d. 5.72). At least 35% of patients were diagnosed as being of paranoid sub-type (not all patients were given a sub-type diagnosis). Control participants were recruited via word of mouth and by posters displayed in hospitals and surgeries.

TABLE 1 ABOUT HERE

Procedure

All participants (patients and controls) were given a battery of neuropsychological tests, administered by an appropriately qualified psychologist. This included the National Adult Reading Test (NART) (Nelson, 1982), and a verbal fluency test for the category of 'animals'. The patient group also underwent a clinical interview with a psychiatrist who administered the Scale for Assessment of Negative Symptoms (SANS) (Andreasen, 1984a), and the Scale for Assessment of Positive Symptoms (SAPS) (Andreasen, 1984b).

For the category fluency test each participant was asked to name as many animals as possible in sixty seconds. Responses were recorded on audio-tape and transcribed to a computer database by a native English speaker. Following transcription of the fluency data, plural names were re-coded using the singular form (e.g. geese became goose), but infant, gender-specific or breed names were not converted to the more generic form (e.g. *puppy*, *bitch*, or *greyhound* were not recoded as *dog*). Statistics relevant to the interpretation of the fluency data, including matrices of inter-item dissimilarity scores, were generated using custom software written by the authors. Additive Tree (*addtree*) hierarchical cluster analyses (Sattath & Tversky, 1977), were subsequently generated from the dissimilarity matrices using Corter's (1982) ADDTREE/P program, whilst MDS representations were created using the SPSS ALSCAL program (Takane, Young, & de Leeuw, 1977). An alpha level of 0.05 was used for all statistical comparisons reported below.

Results

A number of cognitive measures relating to the category fluency task, together with statistical comparisons and effect sizes, are summarized in table 1. The patient group named a total of 703 items in their animal fluency lists of which 31 were repeated items (within lists), generating a mean of 16.80 unique items (item count after excluding repeats). The control group listed 648 animal names in all, of which 15 were repeated items, giving an average of 22.61 unique items. Since patients had lower NART scores than controls, an ANCOVA was performed on these measures of list-length with NART as a covariate,[3] this showed that patients generated significantly fewer total items and unique items than controls independent of their NART score.

Some authors have suggested that semantic memory is particular disintegrated in specific sub-groups of patients with schizophrenia (Paulsen et al., 1996; Rossell et al., 1999;

Sumiyoshi et al., 2001; Sumiyoshi et al., 2005a), therefore we here summarize some specific statistics concerning tests for clinical symptoms. Scores for negative symptoms (on the SANS scale) were in the range 1–18 with a median value or 9, and in the alogia (poverty of speech) sub-scale scores were in the range 0–4 with a median value of 1. 47.5% of patients had a high alogia score, according to the criteria suggested by Sumiyoshi et al. (2005a), of 2 or greater. Scores for positive symptoms (on the SAPS scale) were in the range 0–8 with a median value of 4, and on the delusion sub-scale of this test in the range 0–4 with a median of 1. 45% of patients had a high score on the delusion sub-scale, according to the criteria suggested by Rossell et al. (1999), of 2 or more.

Identifying high frequency items in verbal fluency data

Fluency data for the animal category were first analyzed in terms of production frequencies, that is, the number of participants in each group that named a particular animal. Across both groups, participants generated a total of 206 different animal names. Production frequencies for 39 of these names generated by at least 10% of participants in both patient and control groups were calculated, and found to be strongly correlated (r= 0.87). Nine of the ten most frequently named animals were the same in both groups. These results indicate that patients with schizophrenia are similar to controls in terms of the animal names that typically appear in their category lists.

Twelve animals names—*cat, cow, dog, elephant, giraffe, horse, leopard, lion, monkey, mouse, tiger, sheep*—were named by at least 30% of participants in *both* groups and were therefore selected for investigation using clustering and scaling methods. This follows the practice of earlier studies where analyses of semantic organization in clinical groups have been based on similar sets of high frequency items.[4] This set of twelve high frequency items

is denoted A12 below, while the full set of 206 items generated across both groups is denoted A206.

Items belonging to A12 appeared a total of 257 times in the verbal fluency listings of patients, each naming, on average, 6.43 items. The control group listed A12 items 222 times with an average of 7.93 items per participant. Statistical results relevant to these high frequency items are also summarized in table 1. Since controls generated significantly more target items than patients it was important to consider the impact of this on measures of inter-item dissimilarity (as explained above, this was the motivation for including more patients than controls in the current study). For the sixty-six animal pairs generated by A12, the 40 participants in the patient group produced a total of 802 item co-occurrences with a median number of 11 per animal pair, while the control group generated a total of 809 co-occurrences, median 11.5. Wilcoxon signed rank tests (Table 1) showed that the two groups did not differ significantly in terms of the total number of co-occurrences per animal pair or in the proportion of co-occurrences involving repeated items. Thus we conclude that the two samples were well matched in terms of the number of raw inter-item distances used to compute the dissimilarity matrices.

Comparison of dissimilarity metrics on the basis of frequency distribution

In order to compare the general behavior of the two dissimilarity metrics defined earlier we first examine the frequency distributions that arise when both are applied to the 68 fluency lists generated by the patient and control groups combined (henceforth abbreviated by the group name ALL), and to the 206 animal names listed by either group. Figure 1 shows this comparison both for all elements of the two dissimilarity half-matrixes, $\mathbf{MCF}_{\mathrm{ALL,A206}}$ and $\mathbf{DIS}_{\mathrm{ALL,A206}}$,[5] and for the two underlying measures of normalized inter-item distance, $\hat{D}_{abl}$ and d_{abl}/n_l, on which they are based. The shape of the distributions shown here are consistent

with our mathematical analysis of the two dissimilarity measures (see Supplemental Data Web Page). First, the frequency-based measure of normalized inter-item distance, $\hat{D}_{abl}$, generates an essentially flat distribution with a mean of 0.50. When values of $\hat{D}_{abl}$ are averaged across several lists to calculate the *mcf* metric this generates a mildly humped distribution (a consequence of some regression to the mean) again centered on a value of 0.50. By comparison, the measure based on list-length, d_{abl}/n_l, generates a distribution that is significantly skewed towards the lower end of the range (mean 0.34). In averaging values of d_{abl}/n_l, and weighting these by the inverse of the number of item co-occurrences ($1/T_{gab}$), the *dis* metric generates a distribution that is even more skewed with a mean of just 0.26 and more than half of the data-points falling in the first 20% of the normalized range.

FIGURE 1 ABOUT HERE

Having demonstrated the behavior of the two metrics in relation to the full set of inter-item distances derived from our verbal fluency data, we focus, in most of the remaining analyses, on the set A12 of high frequency animal names and the sixty-six item pairs they give rise to.

Correlational analyses of dissimilarity measures

We next consider some correlational analyses of the performance of our two dissimilarity measures.

First, the elements of the $\mathbf{MCF}_{ALL,A12}$ and $\mathbf{DIS}_{ALL,A12}$ matrices show a correlation of 0.69 for the fluency data of the combined participant group. Thus, although the two metrics are similar, less than half of the variability ($r^2 = 0.48$) of one is captured by the other.

The correlation between the patient and control groups, on either metric, is also worth investigating as it gives an indication of the similarity of the underlying semantic structures in the two groups and of the capacity of the selected metric to distinguish such similarity. The full correlations between the two groups are 0.67 for $\mathbf{DIS}_{\text{PATIENTS,A12}}$ vs. $\mathbf{DIS}_{\text{CONTROLS,A12}}$ and 0.75 for $\mathbf{MCF}_{\text{PATIENTS,A12}}$ vs. $\mathbf{MCF}_{\text{CONTROLS,A12}}$. However, these correlations may reflect other resemblances between the groups than those concerning just semantic distance. In particular, we noted earlier that the *dis* metric varies with frequency of item co-occurrence, hence the observed relationship may partly reflect the high correlation (0.87) between production frequencies in the two groups. A better measure of the similarity in semantic structure, *per se*, is the partial correlation after removing the effect of item frequency.[6] This reduces the between-group correlation to 0.48 for the *dis* metric while it remains high, 0.72, for *mcf*, thus the latter appears to capture more of the similarity in semantic structure between patients and controls.

Cluster analyses of semantic structure using different dissimilarity measures

The above results demonstrate substantial differences between the new dissimilarity measure and the prevailing approach. Past research using the *dis* metric has focused on cluster and scaling analyses as the principle means of identifying differences in semantic memory organization between clinical and control populations, it is therefore appropriate that we compare the two measures in relation to such analyses.

Additive tree analyses were carried out, for each participant group, using the A12 set of high frequency items and the two alternative dissimilarity metrics. This gave rise to the four trees shown in Figure 2 derived from the matrices $\mathbf{DIS}_{\text{PATIENTS,A12}}$, $\mathbf{DIS}_{\text{CONTROLS,A12}}$, $\mathbf{MCF}_{\text{PATIENTS,A12}}$, and $\mathbf{MCF}_{\text{CONTROLS,A12}}$.

Each addtree is composed of *nodes* (horizontal lines in Figure 2) and *arcs* (vertical lines). Items are represented by *external* nodes and are formed into clusters by *internal* nodes; all nodes are joined by arcs. The distinctiveness of a cluster and the degree of dissimilarity within a cluster are both indicated by the *length* of nodes, with the longest lengths indicating the greatest dissimilarity. Thus distinctive clusters will be separated by long internal nodes, while tightly-knit clusters will have short external nodes. For any two items in the tree, their dissimilarity is indicated by the sum of the lengths of the nodes in the path between them (again longer= more dissimilar). Note that the length of arcs has no significance.

FIGURE 2 ABOUT HERE

All of the trees shown in Figure 2 show a general decomposition into two main superordinate clusters[7]—one of 'domestic' animals (i.e. those likely to be found in a home or on a farm) that includes *cat, dog, horse, cow,* and *sheep*, and another of 'wild' animals (such as, in the United Kingdom, might be found in a zoo) that includes *elephant, giraffe, monkey, lion, leopard,* and *tiger*. Beyond this primary decomposition, which is consistent with previous analyses of animal category fluency data in normal populations (see Crowe and Prescott, 2003), there are substantial differences between the trees. Perhaps most noticeably, trees derived from the *mcf* measure have *longer internal nodes and shorter external nodes* than the equivalent trees for the *dis* measure. There is also a similar, though less pronounced, difference between the control and patient groups on both metrics. These differences in proportion are quantified in Figure 3 which shows the ratio of the sum of internal node lengths to the sum of external node lengths across both participant groups and dissimilarity measures. This result indicates that the *mcf* measure generates more distinctive and closely-knit clusters

than the *dis* measure, and that, on either measure, the control group generated more distinctive clusters than patients.

FIGURE 3 ABOUT HERE

The relatively weak similarity structure of the $\mathbf{DIS}_{\text{PATIENTS,A12}}$ addtree is also shown by the absence of sub-clusters (in Figure 2a) reflecting strong inter-item associations such as *cat–dog*, *lion-tiger,* and *elephant–giraffe,* all of which are present in the $\mathbf{MCF}_{\text{PATIENTS,A12}}$ addtree (2c). When the trees generated from the $\mathbf{MCF}_{\text{PATIENTS,A12}}$ and $\mathbf{MCF}_{\text{CONTROLS,A12}}$ matrices are compared (2c vs. 2d) there are just two structural differences. First, *mouse*, which is in the 'domestic' animal cluster for the patient group, appears as a 'wild' animal in the tree generated by control group. Second, the sub-cluster *elephant–giraffe* which is present for the patient group is absent for controls. Neither of these differences is indicative of greater semantic memory disorganization in the patient group.

There are two useful statistical measures of the 'goodness of fit' of an addtree to a matrix of inter-item distances both of which are given in Table 2 for the trees shown in Figure 2. The first is the *stress* value (specifically, Kruskal's stress measure 1) which ranges from 0 (best possible fit) to 1 (worst possible fit) and summarizes the extent to which distances between items in the tree accurately capture the distances provided in the original matrix. High stress values may indicate that a tree is not an adequate representation of the similarity structure of the data. The stress values obtained for the trees in our analysis are in the range 0.040–0.068 indicating a 'good' degree of fit based on Kruskal's (1964) guidelines. The second measure is R^2 which describes the proportion of the variability of the distance data in the matrix that is accounted for by the tree, representations that fit the data well having high R^2 values (i.e. close to 1). For the trees shown in Figure 2, R^2 varies between 0.70 and 0.89,

showing that a fair proportion of the variability is captured in all cases, but that more variability is accounted for in the control group addtrees than in those for the patient group. Note that these goodness of fit statistics do not particularly distinguish between the *dis* and *mcf* metrics.

TABLE 2 ABOUT HERE

MDS analyses of semantic structure

MDS analyses were carried out, for each participant group, using the A12 set of high frequency items and the two dissimilarity metrics. Each analysis used a Euclidean distance model with interval measurement. Two-dimensional solutions were selected, as in previous investigations (Aloia et al., 1996; Moelter et al., 2001; Paulsen et al., 1996; Rossell et al., 1999; Sumiyoshi et al., 2001; Sumiyoshi et al., 2005a), to meet the criterion that the number of stimuli minus one should equal, or be greater than, four times the dimensionality (Kruskal & Wish, 1978). This procedure gave rise to the four maps shown in Figure 4 derived from the $\mathbf{DIS}_{\text{PATIENTS,A12}}$, $\mathbf{DIS}_{\text{CONTROLS,A12}}$, $\mathbf{MCF}_{\text{PATIENTS,A12}}$, and $\mathbf{MCF}_{\text{CONTROLS,A12}}$ matrices. The items in each map are also coded according to aspects of the relevant addtree analysis. Specifically, items belonging to the 'domestic' and 'wild' animal clusters are indicated by filled diamonds and crosses respectively, and items that form a two node sub-cluster in the relevant tree are linked by a dotted line to indicate this association.

The superordinate clusters ('domestic' and 'wild' animals) are not clearly separated in the map generated by the *dis* metric for patient data (4a), although they are better distinguished in that generated by the controls (4b). At the same time, those items that are clustered in the addtree analyses for this metric (e.g. the two item sub-clusters) are often relatively far apart in the corresponding MDS map. Indeed, the organization of the *dis* maps appear, in part, to reflect the strong relationship between this measure and production frequency noted earlier, with high-frequency items tending to be grouped closer to the origin.[8] This influence has the effect of masking inter-item relationships due to semantic distance. By

comparison, the *mcf* metric produced relatively distinct super-ordinate clusters for both patients and controls. Pairs of associated items are also located closer together in these maps consistent with the relatively tightly-grouped sub-clusters seen in the addtrees.

Previous authors have sought to interpret the dimensions of maps of animal fluency data in terms of semantic categories or properties. For instance, Chan et al. (1993) suggested that maps for normal adults will have a principle dimension based on the 'domestic/wild' distinction, and a secondary dimension based on size. Whilst the horizontal dimension in the maps shown in Figure 4 is consistent with a 'domestic/wild' axis, the evidence for a size dimension, along the vertical axis, is comparatively weak. For instance, *giraffe* and *mouse* which would be expected to lie at the extremes of the size scale, are both located in lower half of the map for three of the four analyses.

Goodness of fit statistics for each map are given in Table 2. Stress values are in the range 0.15–0.28 indicating relatively poor 'goodness of fit' according to Kruskal's (1964) guidelines, and compared with the equivalent addtree analyses. The R^2 values again indicate a better fit for control participants than for patients with either dissimilarity measure. Less variance is accounted for by the maps of patient data (R^2= 0.64, 0.60) than the corresponding addtrees (0.70, 0.71).

FIGURE 4 ABOUT HERE

Differences between patients and controls on the category fluency task for animals

The above structural analyses have identified strong similarities in semantic organization in patient and control groups as revealed by the animal category fluency task, suggesting that many of the differences identified in previous studies may be due to the choice of a sub-optimal measure of inter-item distance, or im-balanced sampling of the data from the two subject groups. However, our analyses do confirm one pattern found in most previous studies which is that structural analyses generally show a poorer fit for data derived from patient groups than from controls (see Chan & Ho, 2003 for review). Our addtree

analyses also provide some further indication, in the ratio of internal to external node lengths (*I*/*E*), that the clusters found in patient data are less distinctive. In the following, therefore, we describe some additional investigations intended to clarify how performance in the category fluency task differs between patients with schizophrenia and controls.

Variability in performance of the category fluency task

Hitherto we have focused on analyses using the *mean* normalized inter-item distance between pairs of animal names (the *mcf* metric). However, it is also possible to look at the variability in the inter-item distance measure for each item-pair. An appropriate measure for this is the *standard error* of mean as this statistic adjusts for the number of samples available *(Note: use of the standard error is no longer recommended, see correction on p.54)*. Figure 5 shows the distribution of this measure for the patient and control groups generated for the sixty-six item pairs in A12 group of animal names. The graph shows that the distribution of error-size contains more high values for the patient group than for controls, indeed, the average standard error was found to be approximately 16% larger for patients (0.081. s.d. 0.030) than for controls (0.070, s.d. 0.021). A multivariate analysis of variance of this data showed a significant difference between patient and control groups when the *mcf* measure and its standard error are treated as dependent variables ($F=3.20$, $p=0.04$). Univariate analysis further showed that there was negligible difference in the *mcf* scores ($F=1.31$, $p=0.25$), but a significant difference in the standard error scores ($F= 6.22$, $p= .014$, Cohen's $d= 0.43$) (using a Bonferroni adjusted alpha level of 0.025). We therefore conclude that there is greater variance in the normalized distance measures generated from the patient data.

FIGURE 5 ABOUT HERE

Split-half comparisons

The differences in response variability between patients and controls were further confirmed by split-half correlation comparisons between groups. For this purpose the patient and control groups were first partitioned into two equally sized half-groups (i.e. 2x 20 patients, and 2x 14 controls). In order to control for any between-group differences in these half-samples, three further partitions into two equal-sized sub-groups were then made for each full group, with each partition chosen so as to be maximally dissimilar from all others. This procedure therefore generated four split-half comparisons of the patient and control groups respectively, and sixteen half-group comparisons between patients and controls. For each half-group G, $\mathbf{MCF}_{G,A12}$ was then calculated.[9] The four correlations between matrices for patient half-groups were in the range 0.33–0.50 (median 0.39), compared with 0.57–0.65 (median 0.61) for control half-groups. suggesting that the control group are less variable than the patient group. The sixteen correlations between patient half-groups and control half-groups were in the range 0.40–0.73 (median 0.57) thus the patient half-groups also appear to resemble controls more than they do each other.

Semantic organization of medium and high-frequency items for the patient group

We have shown that the use of an appropriate normalized distance measure, and of balanced data sets, can produce structural analyses for patients with schizophrenia in which the decomposition into semantic categories is similar to that seen in a control group. Since the analyses performed in earlier studies were based on sets of upto 16 animal names, it is worth investigating whether the outcomes obtained here are the particular consequence of analyzing just the twelve most high frequency items (A12) in both groups.

To address this question a further analysis was performed using all the animal names listed by at least 20% of the patient group, with the added constraint that each item should co-

occur with every other at least twice (to ensure that every element of the dissimilarity matrix was computed from at least two measurements). This resulted in a set of 20 animal names (A20) that generated 190 pairs with a median rate of co-occurrence (per item pair) of 7 (range 2–27) in the patient fluency data. The addtree generated for the resulting dissimilarity half-matrix $\mathbf{MCF}_{\text{PATIENTS,A20}}$ is shown in Figure 6. As is clear from the figure, the addition of eight further medium frequency items brings about an increase in the number of clusters but does not reveal significant disintegration of memory organization. The new tree has clusters of 'wild' animals (that includes a sub-cluster of big cats and another of large herbivores), 'domestic' animals, farm animals, and reptiles. Only one item, *monkey*, appears out of place (weakly linked to the reptile cluster), possibly because there are no closely related animals, such as other primates, in the A20 set. Since we included more medium frequency items in this analysis the dissimilarity measures generated from the data are more variable than with the A12 set, this is indicated by the higher average standard error per matrix element of 0.094 (s.d. 0.034).

FIGURE 6 ABOUT HERE

Discussion

The category fluency task is one of a range of methodologies that can be used to gain insight into conceptual memory organization (Henley, 1969; Storm, 1980; Chan et al., 1993; Crowe and Prescott, 2003) and has the particular advantages of being a natural task to perform, a straightforward one to explain to participants, and one where the data can be collected quickly and processed easily. Against these advantages there are a number of drawbacks. For

instance, there is the problem of missing data (subjects will generally not list all the items of interest), the need for multiple fluency lists (since the data is intrinsically one-dimensional), and the presence of noise and of extraneous influences in the data (other factors, beside concept similarity, influence word ordering). To overcome these limitations an appropriate algorithm must be applied to derive estimates of semantic distance, and a reasonably large number of participants needs to be investigated to ensure that low frequency items (and item pairs) are adequately sampled. In this article we have presented methodological advances addressing both of these issues that can be summarised as follows.

Our primary objective was to derive a principled method for estimating inter-item dissimilarity using the raw inter-item distance data that can be obtained from fluency lists. In this regard we have described a new dissimilarity measure, the *mcf* metric, that improves on that used in previous studies in several ways. First, the method for controlling for differences in list-length is derived from an analysis of the relevant frequency distribution and thus has a clear rationale. Second, the resulting normalized distance measure is evenly distributed between 0 and 1 and therefore does not introduce any systematic distortion. Third, estimates from multiple participants are combined by simple averaging and without the introduction of confounding influences related to item production frequency. When the *mcf* metric is used to generate cluster and map analyses of fluency data the results show clear evidence of semantic structure with clusters that are distinctive and closely knit. The extent to which this improves over earlier approaches is shown perhaps most clearly in Figure 3 where we demonstrated that a measure of the strength of clustering in an addtree—the ratio of internal to external node length—is much higher (indicating stronger clusters) for the *mcf* metric than for the prevailing *dis* metric.

With regard to the issue of sample size we have also made a methodological innovation relative to earlier research. Specifically, we have a identified a likely confound in

the results from some previous studies arising from the over-sampling of control fluency data compared to that of patients, and due to the reduced fluency output of the latter group. To address this issue, we have analyzed data from patient and control groups where the sample size was selected so that the number of item-pairs (rather than the number of participants) was balanced across groups. It might be objected that this approach simply replaces one uncontrolled difference between the two participant groups with another—the number of items pairs might now be balanced, but group size is now no longer matched. With respect to this issue we would point out the following. First, a number of the statistical analyses we have performed, e.g. the correlations and analysis of variance, make comparisons between the mean normalised distance measures (and their standard errors) for item-pairs across groups, and do not combine this data within a group. Thus for these analyses, which show that patient semantic structures are more variable, the data is appropriately controlled since the average number of samples per item-pair is the same in both groups. Second, whilst the structural analysis do bring together the distance measures for item pairs within a group—and hence the difference in group size could have some impact—it should be noted that our methodology results in a more heterogeneous sample of patients than of controls. This greater variability should work against the semantic structures generated for the patient group showing as much distinct structure as for the controls. That the structural analyses generated for patient groups shows similar structure to that of controls therefore cannot be as the result of our sampling methodology.

The results of our analyses using this balanced data set and the new dissimilarity metric suggest that, overall, in relation to data from the category fluency task, the category 'animals' does not appear to be systematically disorganized in patients with schizophrenia as compared to controls. In the following we consider several specific ways in which memory

for this category might have been distorted and compare the results of the current study with those previously reported.

Organization of the animal category into 'wild' and 'domestic' animals

Perhaps the strongest finding from previous structural analyses of animal fluency data (see Crowe and Prescott 2003, for review) is that subjects, of all ages, tend to differentiate 'wild' animals, such as *elephant, giraffe, lion*, and *tiger*, from domestic ones such as *dog, cat, horse*, and *cow*. This 'wild/domestic' distinction is usually identified with one of the main axes in scaling analyses of animal semantic space and in cluster analyses by the presence of two or more large and distinct clusters in which wild and domestic animals do not mix. With the possible exception of Rossell et al. (1999), previous studies comparing schizophrenic and control groups have found strong evidence for this distinction in the fluency lists generated by controls (Aloia et al., 1996; Moelter et al., 2001; Paulsen et al., 1996; Sumiyoshi et al., 2001; Sumiyoshi et al., 2005a). Most analyses of patient data, however, suggests that this decomposition is either partially disintegrated (Moelter et al., 2001; Sumiyoshi et al., 2005a) or almost entirely absent (Aloia et al., 1996; Sumiyoshi et al., 2001). In the current study all analyses of both control and patient data showed a clear split between wild and domestic animals. The clustering, however, was much less distinctive when generated using the *dis* metric than with the *mcf* metric. Relatively weak clusters generated by the *dis* metric will be more easily disrupted by the increased variability in patient data relative to controls, and this may explain why the wild/domestic distinction appears to be disintegrated in some studies. Interestingly, Paulsen et al. (1996) who conducted one of the few studies to test substantially more patients than controls, also found a fairly clear decomposition along a wild/domestic axis in their patient data.

When medium frequency animal names are analyzed together with high frequency ones it becomes clear that a principal determinant of memory organization for this category is the environmental context in which different animals are commonly encountered (Storm, 1980; Crowe and Prescott, 2003). Thus, typically, there are separate clusters of animals found in the home, on the farm, or in the zoo. A similar pattern is shown in figure 6 where we analyzed patient data for twenty animal names. This result suggests that, on average, memory organization in schizophrenia is also similar to the normal adult population for less prototypical members of the animal category.

Associations between strongly-related items

A second robust finding from the psychological literature is the presence of strong associative links between specific animal pairs such as *cat–dog*, *lion-tiger,* and *horse–cow* (e.g. Grube and Hasselhorn, 1996). Such associations are most easily identified in cluster analyses where two items with relatively short external nodes are linked to a larger cluster, or the main tree, by a relatively long internal node (as, for instance, for *cat-dog* in figure 2d). In scaling analyses strong associations may appear as items that are in close proximity to one another. One of the most striking findings of the study by Aloia et al. (1996) was the absence of such links in their cluster analysis of patient data, whereas strong associations between several stereotypical animal pairs were clearly evident in their equivalent analysis of control data. Other studies, however, have not shown such a prominent difference—strongly associated pairs appear in the cluster analyses of patient data performed by Moelter et al. (2001) and of non-deluded patients tested by Rossell et al. (1999). In the current study a fairly dramatic difference can be seen between the *mcf* and *dis* metrics (see figure 2) whereby the strong associations identified by the *mcf* metric are either absent or significantly weaker in the analyses generated using the *dis* metric. This finding suggests that previous studies, all of

which have used the latter metric, may have significantly underestimated the presence of such associations in patient data.

Organization of the animals category according to size

Since Henley (1969), various authors (e.g. Storm, 1980) have suggested that an important dimension along which semantic memory for the animal category is organized relates to physical size. In studies involving patients with schizophrenia, however, this 'size' dimension is clearly shown in only one scaling analysis of control data (Paulsen et al., 1996) and is either missing altogether (Moelter et al., 2001; Rossell et al., 1999; Sumiyoshi et al., 2005a), or relatively weak (and therefore open to alternative interpretations) in others (Aloia et al., 1996; Sumiyoshi et al., 2001). Structural organization according to physical size was not evident in any of the scaling analyses reported in the current article. Given this mixed evidence for a size dimension in control groups, the absence of an equivalent dimension in scaling analyses of patient data does not provide strong support for the hypothesis of a deterioration of semantic organization in schizophrenia. There are two further reasons to be sceptical about arguments for semantic disorganization based on the absence of a size dimension in scaling analyses. First, as Storms, Dirikx, Saerens, Verstraeten, and De Deyn (2003) point out, MDS is an algorithm that finds the best-fit mapping from a complex data-set (here the matrix of inter-item distances) into a low-dimensional space (here a 2D plane), and hence there is no reason to expect that the dimensions of the resulting projection will have any straight-forward interpretation. Second, as was reported above, the MDS solutions obtained here tend to have poor 'goodness-of-fit' statistics as measured against the widely accepted Kruskal (1964) guidelines, and when compared to the equivalent addtree analyses[10] (see table 2). These results indicate that two-dimensional spatial models may be less

appropriate than cluster-based models as a means for representing concept similarity data for the category animals (see also Rossell et al. 1999).

Greater variability in fluency data from patients with schizophrenia

The current study presents several findings that are consistent with the wider literature on impaired performance by patients with schizophrenia on the category fluency task. First, in common with many previous studies (see Bokat & Goldberg, 2003 for review), the patients tested here generated significantly fewer items in their category lists than controls after allowing for differences in verbal intelligence. Second, when the dissimilarity estimates (elements of the dissimilarity matrix) obtained from patients and controls were compared the former were shown to have greater variability (higher mean standard error). This difference cannot be explained as a consequence of patients generating fewer items in their lists as patient and control group sizes were specifically chosen such that, on average, the number of participants contributing to each element of the dissimilarity matrix did not differ between the two groups. Patients were also found to be less internally consistent than controls on split-half correlation measures. Indeed, the dissimilarity matrices for patient sub-groups had higher correlations with those for control sub-groups than with each other. This pattern, which matches previous findings by Elvevåg and Storms (2003), suggests that whilst the average category structure may be similar between patients and controls (as shown in our structural analyses), in patients there is less consistency across individuals.

There has been considerable debate in the literature as to whether the impaired performance of patients with schizophrenia on the category fluency task is best understood as arising from a specific problem in memory organization, or from a more general cognitive impairment leading to difficulties in accessing or searching a lexical store that is more-or-less intact (see e.g. Bokat & Goldberg, 2003; Elvevåg et al., 2001; Giovannetti et al., 2003; Green

et al., 2004; Vinogradov et al., 2003). Evidence favouring the memory impairment hypothesis includes a recent meta-analysis showing that patients with schizophrenia are disproportionately impaired in category fluency relative to letter fluency (Bokat & Goldberg, 2003). Since letter fluency is thought to be less reliant on semantic processing than category fluency, this result suggests that performance on the latter is due, at least in part, to disorganization of semantic memory structure in addition to any general deficit in attention or retrieval that could be expected to effect both tasks. Structural analyses of data from the category fluency task can contribute to this ongoing debate by helping to characterise the ways in which patient fluency data deviate from that of controls over and above these basic differences in word production. For instance, some studies have claimed that analysis of schizophrenic fluency lists reveals unusual or bizarre structure (Paulsen et al., 1996; Sumiyoshi et al., 2005a) that would be difficult to explain as anything other than a distortion in the underlying knowledge representation. In our study we have shown that patient data is more variable than that of controls after controlling for differences in total fluency output, but we have not found any evidence to support stronger claims of systematic semantic memory disorganization in schizophrenia.

As pointed out by Elvevåg and Storms (2003) and Storms et al (2003), cross-sectional analyses, such as those presented here, that can reveal common patterns of organization or distortion within a group, are less informative where there are significant, but unrelated, differences between individuals. Thus there are two possible explanations for the greater variability seen in our patient data that are not distinguished by the current study. The first possibility is that there are no interesting underlying differences in memory organization from one patient to the next. The observed variability might then be best characterised as due to increased 'noise' in the semantic networks of individuals with schizophrenia, or in the cognitive processes that sample these networks during performance of the fluency task. Such

an outcome would be consistent with evidence that thought processes in schizophrenia may be effected by episodes of chaotic activity that disrupt normal chains of thought (Paulus & Braff, 2003). The second possibility is that patient memory is not intrinsically noisier but simply more eccentric or more idiosyncratic than in normal controls, in other words, that distortions in semantic memory may show an interesting pattern in relation to each individual's clinical condition or life history. Further investigation of these issues will require the use of different methodologies from the group analyses provided here using, for instance, longitudinal studies.

Semantic organization in patient sub-groups

Some authors have suggested that semantic memory is particularly disintegrated in specific sub-groups of patients with schizophrenia. First, Paulsen et al. (1996) have suggested that memory for the animal category is most disorganized in patients of the paranoid sub-type (compared with other sub-types) and in patients in whom schizophrenia was first diagnosed before the age of 45. Second, Rossell et al. (1999) have suggested that there is more memory disorganization in deluded patients (i.e. those with SAPS delusion scores of 2 or greater) compared to the non-deluded. Finally, Sumiyoshi and co-workers (2001, 2005a) have made similar claims for patients with very early onset schizophrenia (first diagnosed before age 20), low vocabulary scores (WAIS-R 7 or less), or with high levels of alogia (SANS alogia scores of 2 or greater). The wide variety of sub-groups implicated by these studies raises doubts as to whether the semantic deficit in schizophrenia is specific to any one sub-group but is consistent with the possibility that memory impairment is greater in patients affected from an early age or exhibiting more profound psychiatric symptoms.

As noted in the methods section, the schizophrenic group analyzed in the current study included a substantial proportion of patients belong to most of the sub-groups listed

above, thus it seems unlikely that the findings we have reported are due to the failure to sample critical sub-populations that are most prone to semantic memory deficit. One limiting factor in the current study is that the patient sample was entirely male. However, since there is little evidence of gender-specific differences in verbal fluency in people with schizophrenia (see Paulsen et al. 1996), it seems unlikely that this alone can account for the differences between the current findings and those of earlier studies.

One of the benefits of investigating patient sub-groups is that these are likely to be more homogenous in terms of symptoms (and potentially their underlying causes) than the schizophrenic population as a whole (Sumiyoshi et al., 2005a). Thus we believe that further investigations of specific patient sub-groups would be worthwhile to identify whether the findings of these earlier studies can be replicated using the new dissimilarity metric proposed here.

Effects of medication

There is good evidence to suggest that atypical antipsychotic medications may have an ameliorative effect on some of the cognitive deficits seen in schizophrenia (see, e.g. Harvey, Green, McGurk, & Meltzer, 2003). In relation to the category fluency task, several studies have reported an increase in the number of words generated by patients after treatment with the atypical drugs used in the current study (Buchanan, Holstein, & Breier, 1994; Lee, Thompson, & Meltzer, 1994; McGurk, Lee, Jayathilake, & Meltzer, 2004; Meltzer & McGurk, 1999; Harvey *et al*., 2003; Stip *et al*., 2003). It should be noted, however, that the effect sizes in these studies are relatively small[11], do not restore function to normal levels, and that a number of studies involving olanzapine (the medication used by almost half of the patients investigated here) have not shown a significant improvement in fluency output (Cuesta, Peralta, & Zarzuela, 2001; Purdon *et al*., 2000). Further, in the current study, none

of the patients were treated with ziprasidone, which is possibly the most potent atypical with regard to impact on cognitive dysfunction (Harvey et al., 2004). A recent study by Sumiyoshi et al. (2005b) used MDS and hierarchical clustering analyses of animal fluency data to investigate the possible effects of treatment with atypical drugs on semantic memory organisation. This work showed some evidence for better organisation following six weeks of treatment with atypicals (either olanzapine or ziprasidone) in that it became possible to discern a wild-domestic split in the resulting MDS analyses that was not evident at baseline. However, for both drug treatments the resulting cluster analyses still showed substantial disintegration of these super-ordinate categories. Since Sumiyoshi et al. used the prevailing *dis* metric, the results obtained in that study will have been affected by the distortion that we have shown to arise from the use of that measure. Since the majority (77.5%) of the patients in the current study were on atypical medication, differences in drug treatment compared to earlier studies are a possible factor in explaining the presence of a wild/domestic decomposition in our analyses generated using the *dis* metric. The structural differences we have described between analyses generated using the *dis* and *mcf* metrics are, however, clearly independent of any medication effects.

The use of the category fluency task as a tool for investigating conceptual structure

Good estimates of human conceptual memory structure are not easy to obtain. Psychologists have traditionally approached this problem by adopting concept similarity as an indicator of how concepts are organized in memory (see, e.g. Medin, 1989), and have then set about obtaining multiple measure of concept similarity between different item pairs. For instance, one option is to present participants with many different triads—e.g. [*cat cow dog*]—asking them to select the two most similar items in each case (e.g. Henley, 1969; Chan et al, 1995); another is to collect priming data since concepts that are good primes for one another can also

be assumed to be semantically related (e.g. Neely, 1991); a third is to collect data on word associations (e.g. Henley, 1969). Whilst these approaches differ in a number of ways, they all have in common that individuals must generate a large number of responses in order that a full matrix of item-to-item comparisons can be generated. By comparison, the category fluency task is quick and easy to administer, and, since it is frequently used as a diagnostic tool, fluency data is readily available for many clinical groups without the need for special-purpose testing. This ease of use should not be allowed to mislead us, however, as to the utility of this test for understanding conceptual memory organisation. By pooling data across a sufficiently large pool of subjects, as we have done here, the fluency task can provide a quick, if somewhat fuzzy, snapshot of the typical pattern of concept similarity. However, if the goal is to obtain a picture of the semantic organisation on an individual basis, or to explore differences within, say, a specific clinical population, data from the fluency task, particularly in its shortened diagnostic form, cannot substitute for more comprehensive testing using something like triadic comparison.

Having obtained concept-similarity data from a fluency, triad, priming, or word association study a matrix will have been generated consisting of a very large number of pair-wise comparisons. The use of some data reduction algorithm is then essential to make sense of all these measurements. Structural analysis tools such as the multidimensional scaling (MDS) and hierarchical cluster analyses used here have been developed specifically for this purpose so it is natural that they should applied to this problem (see also Chan & Ho, 2003). However, it is important to note that these tools are primarily intended for finding structure in data rather than for hypothesis testing. As such there are few useful tests for quantifying the extent to which one structural outcome is similar to another, comparisons thus tend to be made qualitatively. Where possible, therefore, other statistical techniques should be applied to determine whether the differences between the matrices obtained from the chosen

participant groups achieve statistical significance (see Elvevâg and Storms, 2003, for more discussion of this issue). For instance, in the current article we have used multivariate analysis of variance to examine differences in variability between participant groups. In addition, studies should be devised that test for hypothesized memory distortions or deficits in a more direct way than is possible by extrapolating patterns of semantic memory organisation from fluency data.

Conclusion

This article has identified significant methodological limitations in past studies that have used cluster and scaling analyses of category fluency data to investigate conceptual memory organization in patients with schizophrenia. We have shown how these limitations can be overcome and have presented new analyses suggesting that, in relation to performance on this task, the extent of memory disorganization in chronic schizophrenia may have been over-estimated. Specifically, we have shown that the use of an improved dissimilarity metric reveals, in patient data, an average pattern of semantic relationships between animal concepts that shares many of the organizational characteristics observed in control data. Whilst our results do demonstrate greater variability within the clinical group, they do not indicate any systematic pattern of memory distortion across patients. Overall, this evidence is consistent with some disintegration of memory storage in schizophrenia but suggests that such deterioration may not be as marked as previously supposed.

References

Aloia, M. S., Gourovitch, M. L., Weinberger, D. R., & Goldberg, T. E. (1996). An investication of semantic space in patients with schizophrenia. *Journal of the International Neuropsychological Society, 2*, 267-273.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders - fourth edition (DSM-IV)*. Washington D.C.: Author.

Andreasen, N. C. (1984a). *Scale for the assessment of negative symptoms (SANS)*. Iowa City, IA: University of Iowa.

Andreasen, N. C. (1984b). *Scale for the assessment of positive symptoms (SAPS)*. Iowa City, IA: University of Iowa.

Bokat, C. E., & Goldberg, T. E. (2003). Letter and category fluency in schizophrenic patients: A meta-analysis. *Schizophrenia Research, 64*(1), 73-78.

Buchanan, R. W., Holstein, C., & Breier, A. (1994). The comparative efficacy and long-term effect of clozapine treatment on neuropsychological test performance. *Biological Psychiatry*, *36*(11), 717-725.

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 Alzheimers- disease. *Journal of Cognitive Neuroscience, 5*(2), 254-261.

Chan, A. S., Butters, N., Salmon, D. P., Johnson, S. A., Paulsen, J. S., & Swenson, M. R. (1995). Comparison of the semantic networks in patients with dementia and amnesia. *Neuropsychology, 9*(2), 177-186.

Chan, A. S., & Ho, Y. C. (2003). Things aren't as bad as they seem: A comment on Storms et al. (2003). *Neuropsychology, 17*(2), 302-305; discussion 323-309.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Corter, J. E. (1982). Addtree/P: A pascal program for fitting additive trees based on sattath & tversky's addtree program. *Behavior Res. Methods and Instrumentation, 14*, 353-354.

Crowe, S. J., & Prescott, T. J. (2003). Continuity and change in the development of category structure: Insights from the semantic fluency task. *International Journal of Behavioral Development, 27*(5), 467-479.

Cuesta, M. J., Peralta, V., & Zarzuela, A. (2001). Effects of olanzapine and other antipsychotics on cognitive function in chronic schizophrenia: A longitudinal study. *Schizophrenia Research, 48*(1), 17-28.

Elvevåg, B., Weinstock, D. M., Akil, M., Kleinman, J. E., & Goldberg, T. E. (2001). A comparison of verbal fluency tasks in schizophrenic patients and normal controls. *Schizophrenia Research, 51*(2-3), 119-126.

Elvevåg, B., & Storms, G. (2003). Scaling and clustering in the study of semantic disruptions in patients with schizophrenia: A re-evaluation. *Schizophrenia Research, 63*(3), 237-246.

Giovannetti, T., Goldstein, R. Z., Schullery, M., Barr, W. B., & Bilder, R. M. (2003). Category fluency in first-episode schizophrenia. *J Int Neuropsychol Soc, 9*(3), 384-393.

Green, N. E., Done, D. J., Anthony, S. H., McKenna, P. J., & Ochocki, M. (2004). Can disorganisation of semantic memory account for the abnormalities of thought in schizophrenia-a controlled experimental study. *Schizophrenia Res., 70*(2-3), 233-240.

Gruenewald, P. J., & Lockhead, G. R. (1980). The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory, 6*(3), 225-241.

Harvey, P. D., Green, M. F., McGurk, S. R., & Meltzer, H. Y. (2003). Changes in cognitive functioning with risperidone and olanzapine treatment: A large-scale, double-blind, randomized study. *Psychopharmacology* (Berl), *169*(3-4), 404-411.

Harvey, P. D., Meltzer, H., Simpson, G. M., Potkin, S. G., Loebel, A., Siu, C., et al. (2004). Improvement in cognitive function following a switch to ziprasidone from conventional antipsychotics, olanzapine, or risperidone in outpatients with schizophrenia. *Schizophrenia Research, 66*(2-3), 101-113.

Henley, N. M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behaviour, 8*, 176-184.

Kruskal, J. B. (1964). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29*, 1-27.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverley Hills, CA: Sage Publications.

Lee, M. A., Thompson, P. A., & Meltzer, H. Y. (1994). Effects of clozapine on cognitive function in schizophrenia. *Journal of Clinical Psychiatry*, *55* Suppl B, 82-87.

McGurk, S. R., Lee, M. A., Jayathilake, K., & Meltzer, H. Y. (2004). Cognitive effects of olanzapine treatment in schizophrenia. *Medscape General Medicine*, *6*(2), 27.

Medin, D. L. (1989). Concepts and conceptual structure. *American Psychologist, 44*(12), 1469-1481.

Meltzer, H. Y., & McGurk, S. R. (1999). The effects of clozapine, risperidone, and olanzapine on cognitive function in schizophrenia. *Schizophrenia Bulletin*, 25(2), 233-255.

Moelter, S. T., Hill, S. K., Ragland, J. D., Lunardelli, A., Gur, R. C., Gur, R. E., et al. (2001). Controlled and automatic processing during animal word list generation in schizophrenia. *Neuropsychology, 15*(4), 502-509.

Neely, J. H. (Ed.). (1991). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner & G.W. Humphreys (editors)

Basic Processes in Reading: Visual Word Recognition. Hillsdale, N.J: Lawrence Erlbaum Associates.

Nelson, H. E. (1982). *National adult reading test (nart): Test manual*. Windsor, UK: NTER.Nelson.

Paulsen, J. S., Romero, R., Chan, A., Davis, A. V., Heaton, R. K., & Jeste, D. V. (1996). Impairment of the semantic network in schizophrenia. *Psychiatry Research, 63*(2-3), 109-121.

Paulus, M. P., & Braff, D. L. (2003). Chaos and schizophrenia: Does the method fit the madness? *Biological Psychiatry*, *53,* 3–11.

Purdon, S. E., Jones, B. D., Stip, E., Labelle, A., Addington, D., David, S. R., et al. (2000). Neuropsychological change in early phase schizophrenia during 12 months of treatment with olanzapine, risperidone, or haloperidol. The canadian collaborative group for research in schizophrenia. *Archives of General Psychiatry*, *57*(3), 249-258.

Rossell, S. L., Rabe-Hesketh, S. S., Shapleske, J. S., & David, A. S. (1999). Is semantic fluency differentially impaired in schizophrenic patients with delusions? *Journal of Clinical and Experimental Neuropsychology, 21*(5), 629-642.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika, 42*(3), 319-345.

Stip, E., Remington, G. J., Dursun, S. M., Reiss, J. P., Rotstein, E., MacEwan, G. W., et al. (2003). A Canadian multicenter trial assessing memory and executive functions in patients with schizophrenia spectrum disorders treated with olanzapine. *Journal of Clinical Psychopharmacology*, *23*(4), 400-404.

Storm, C. (1980). The semantic structure of animal terms - a developmental-study. *International Journal of Behavioral Development, 3*(4), 381-407.

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*(2), 289-301.

Sumiyoshi, C., Matsui, M., Sumiyoshi, T., Yamashita, I., Sumiyoshi, S., & Kurachi, M. (2001). Semantic structure in schizophrenia as assessed by the category fluency test: Effect of verbal intelligence and age of onset. *Psychiatry Research, 105*(3), 187-199.

Sumiyoshi, C., Sumiyoshi, T., Nohara, S., Yamashita, I., Matsui, M., Kurachi, M., et al. (2005a). Disorganization of semantic memory underlies alogia in schizophrenia: An analysis of verbal fluency performance in Japanese subjects. *Schizophrenia Research, 74*(1), 91-100.

Sumiyoshi, C., Sumiyoshi, T., Roy, A., Jayathilake, K., & Meltzer, H. Y. (2005b). Atypical antipsychotic drugs and organization of long-term semantic memory: Multidimensional scaling and cluster analyses of category fluency performance in schizophrenia. *International Journal Neuropsychopharmacology*, (Epub ahead of print) doi:10.1017/S1461145705006310.

Takane, Y., Young, F. W., & de Leeuw, J. (1977). Nonmetric individual differences multidimensional scaling: An alternating least squares method with optimal scaling features. *Psychometrika, 42*, 7-67.

Vinogradov, S., Kirkland, J., Poole, J. H., Drexler, M., Ober, B. A., & Shenaut, G. K. (2003). Both processing speed and semantic memory organization predict verbal fluency in schizophrenia. *Schizophrenia Research, 59*(2-3), 269-275.

Wixted, J. T., & Rohrer, D. (1994). Analyzing the dynamics of free-recall - an integrative review of the empirical literature. *Psychonomic Bulletin & Review, 1*(1), 89-106.

Author Note

Tony J. Prescott*

Adaptive Behaviour Research Group, Department of Psychology, University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom.

Lisa D. Newton, Nusrat U. Mir, Peter W. R. Woodruff, and Randolph W. Parks

Sheffield Cognition And Neuroimaging Laboratory (SCANLab), Academic Clinical Psychiatry, University of Sheffield School of Medicine, The Longley Centre, Norwood Grange Drive, Sheffield S5 7JT, United Kingdom.

This research was supported by the Sheffield Care Trust NHS, and by a Margaret Temple Award from the British Medical Association.

*Please address correspondence to this author (email: t.j.prescott@sheffield.ac.uk).

Footnotes

[1]In a list of length *n* there are $n(n-1)/2$ item pairs. For each group in the Aloia et al. study we can insert the average number of target items into this equation, and then multiply by the number of participants, to give an estimate of the total number of item pairs. This calculations suggests approximately 375 raw distance measures for the patient group and 672 for the control group. Similar calculations are not possible for other studies since they do not report the number of target items generated by each group.

[2]It is worth noting that Chan et al.'s (1993) article on patients with Alzheimer's disease provided similar statistics and also tested twice as many patients as controls in order to formulate balanced data-sets.

[3]Since NART correlated with years of education (r= 0.54) only the former was used in the ANCOVA. List length increased significantly with NART score independent of condition ($F_{(1,62)}= 5.11$, p= 0.03, for unique items), but the interaction between condition and NART was not significant ($F_{(1,61)}= 0.88$, p= 0.35).

[4]Previous studies have looked at sets of between 11 and 17 animal names that appeared with high frequency in their category fluency data. The actual items used have varied somewhat between studies, partly as the result of differences in local culture and ecology that affect the frequency with which specific animal names appear. However, 7 out of the 12 animal names used here—*cat, cow, dog, elephant, giraffe, lion, tiger*—were also investigated in all previous studies.

[5]In order to generate a normalised measure with range 0–1 the scaling factor N_G was omitted from Equation 4 in the calculation of the *dis* metric. This change has no effect on the shape of the distribution of dissimilarity scores.

[6]Note that partial correlation is not ideal since it controls for linear effects and we know from equation 4 that the relationship between item co-occurrence and the *dis* measure

is non-linear. Thus the reported correlations may still over-estimate the strength of the relationship between groups independent of item frequency.

[7]The placement of the root node of an addtree is not constrained by the algorithm that determines inter-node distances and is therefore usually chosen so as to minimise the overall depth of the tree. This was the case for the two analyses of patient data shown in Figure 2, however, for the control group the root node was selected so as to best distinguish the two main clusters rather than to minimize tree-depth. Note that this has no effect on the way that items are grouped within the tree.

[8]For *dis* analyses production frequency accounts for 83–84% of the variance in the radial distance from the origin (plot center), for *mcf* analyses the equivalent figure was 25–52% of variance.

[9]The number of item co-occurrences contributing to the calculation of each matrix was similar (median 401 for patients and 405 for controls), thus the half-group analyses are appropriately balanced in terms of the number of data points used.

[10]On the basis of poor 'goodness-of-fit' statistics, Storms et al. (2003) have argued that the patient data presented in some previous studies "did not result in significantly better fits […] than random data" (Storms et al. 2003, p. 293). Whilst the stress values for the MDS analyses presented here fall within the range identified as problematic by these authors, those for the addtree analyses are much better and are therefore not vulnerable to this criticism.

[11] For instance, in a large, double-blind study, investigating changes in cognitive function following treatment with atypicals, Harvey et al. (2003) found an effect size (Cohen's d), for increase in category fluency output, of just 0.10 after eight weeks treatment with olanzapine. This is a relatively small effect (Cohen, 1988) and corresponds to an overlap between distributions (baseline vs. post-treatment) of ~93%.

Table 1

Demographic and Cognitive Variables Relating to the Patient and Control Groups

Measure	Patients	Controls	test	p	d
Number of participants	40	28			
Age in years (M)	32.80 (8.40)	29.46 (7.45)	$t_{(66)}$= 1.67	0.10	0.42
Years in education (Mdn)	11 (11–16)	13 (11–18)	U= 321	<0.01	
NART (M)	99.58 (12.71)	111.52 (11.54)	$t_{(63)}$= 3.87	<0.01	0.98
CFT Animals:					
Total responses (M)	17.58 (5.38)	23.14 (7.23)	$F_{(1,62)}$= 6.26	0.02	0.91
Unique items (M)	16.80 (5.42)	22.61 (7.17)	$F_{(1,62)}$= 7.12	0.01	0.97
Repeated items (f)	4.6%	2.4%	U= 531	0.69	
Target items (A12, M)	6.43 (2.32)	7.93 (1.72)	$t_{(66)}$= 2.93	<0.01	0.73
Item pairs:					
Co-occurrences (Mdn)	11 (3–27)	11.5 (4-26)	z= 0.22	0.83	
Repeat co-occurrences (f)	11.9%	8.4%	z= 1.34	0.18	

Note: Figures in brackets indicate the standard deviation of the mean (M) or, where the median (Mdn) is reported, the range; frequencies (f) of repetitions are reported as percentages. The final column is the measure of effect size (Cohen's d). For total responses and unique items statistical results are after adjusting for differences in NART score (three participants—two patients, one control—failed to complete this test).

Table 2

Measures of 'goodness of fit' for the addtree and MDS analyses

	Addtree		MDS	
Dissimilarity matrix	Stress	R^2	Stress	R^2
$\mathbf{DIS}_{PATIENTS,A12}$	0.040	0.70	0.28	0.64
$\mathbf{DIS}_{CONTROLS,A12}$	0.060	0.89	0.15	0.88
$\mathbf{MCF}_{PATIENTS,A12}$	0.055	0.71	0.24	0.60
$\mathbf{MCF}_{CONTROLS,A12}$	0.068	0.87	0.15	0.87

Figure Captions

Figure 1. Frequency distribution of all elements of the dissimilarity half-matrices $\mathbf{MCF}_{\mathrm{ALL,A206}}$ and $\mathbf{DIS}_{\mathrm{ALL,A206}}$, and of the two underlying measures of normalized inter-item distance, $\hat{D}_{abl}$ and d_{abl}/n_l. The graph demonstrates that metrics based on cumulative frequency ($\hat{D}_{abl}$ and *mcf*) generate a more-evenly distributed set of measurements than those based on division by list-length (d_{abl}/n_l and *dis)*.

Figure 2. Addtree analyses of verbal fluency data, for the A12 set of high frequency animal names, from *patient* (trees a, b) and *control* (trees c, d) groups, using the alternative dissimilarity metrics *dis* (trees a, c) and *mcf* (trees b, d).

Figure 3. The ratio of the sum of internal node lengths to the sum of external node lengths (I/E) compared across dissimilarity measures and patient and control groups. The *mcf* measure generates trees with more distinctive sub-clusters than the *dis* measure; there is a similar, but less pronounced, difference between the control and patient groups.

Figure 4. MDS analyses of verbal fluency data, for the A12 set of high frequency animal names, from *patient* (maps a, b) and *control* (maps c, d) groups, using the alternative dissimilarity metrics *dis* (maps a, c) and *mcf* (maps b, d).

Figure 5. Frequency distribution, over sixty-six item pairs, of the *standard error* of the *mcf(G,a,b)* measure for patient and control groups. The graph shows the distance estimates obtained from the patient group are more variable than those of controls.

Figure 6. Addtree analyses for items named by at least 20% of the patient group (and co-occurring with each other item at least twice) based on the matrix $\mathbf{MCF}_{PATIENTS,A20}$. Goodness of fit measures (stress = 0.06, R^2= 0.59) were similar to those reported for the A12 analyses.

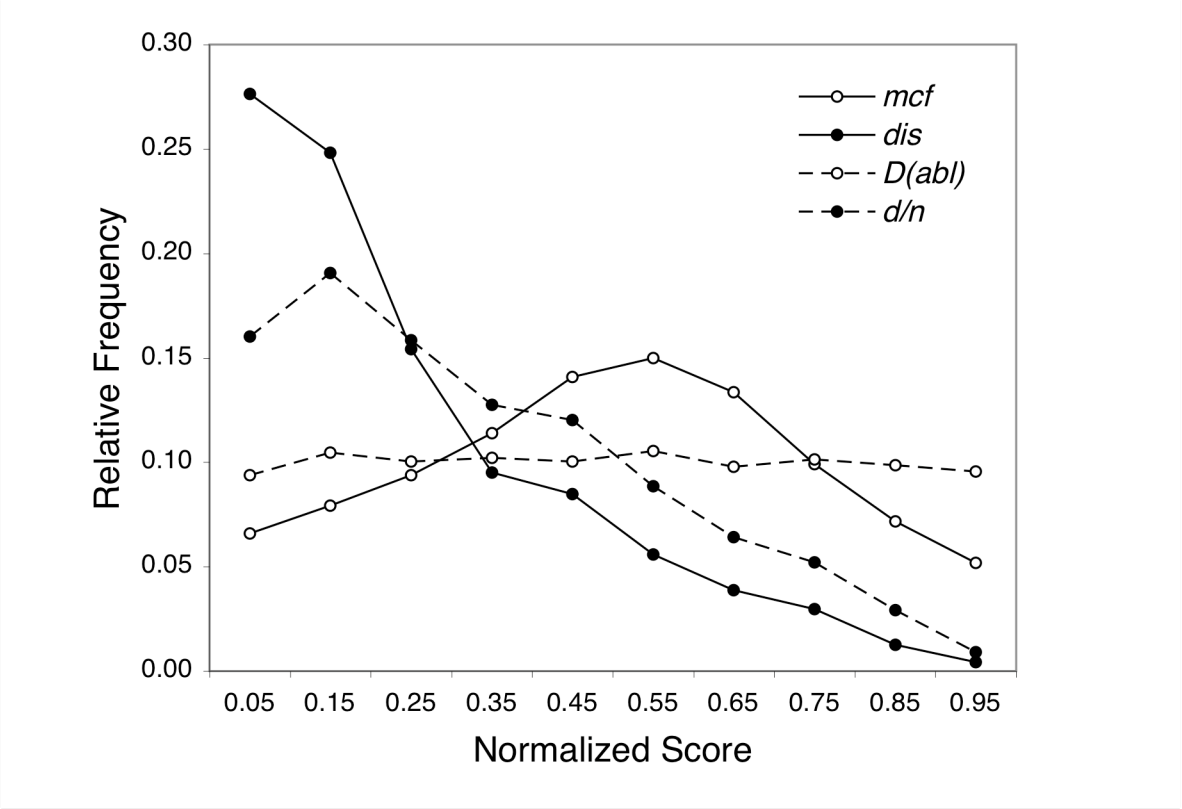

0.30
0.25
0.20
0.15
0.10
0.05
0.00
Relative Frequency
0.05
0.15
0.25
0.35
0.45
0.55
0.65
0.75
0.85
0.95
Normalized Score
mcf
dis
D(abl)
d/n

Figure 1

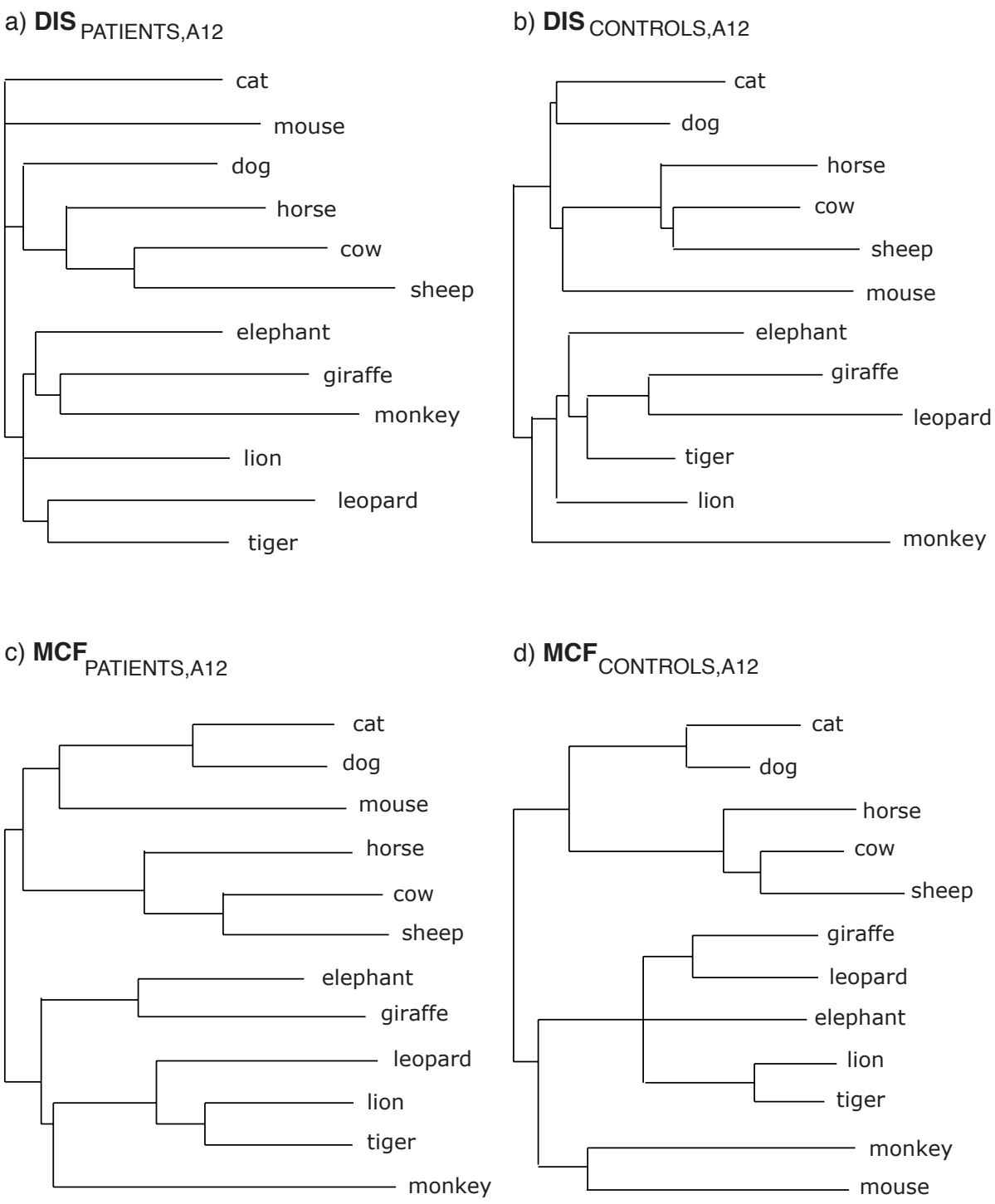

a) DIS PATIENTS,A12
cat
mouse
dog
horse
cow
sheep
elephant
giraffe
monkey
lion
leopard
tiger
b) DIS CONTROLS,A12
cat
dog
horse
cow
sheep
mouse
elephant
giraffe
leopard
tiger
lion
monkey
c) MCF PATIENTS,A12
cat
dog
mouse
horse
cow
sheep
elephant
giraffe
leopard
lion
tiger
monkey
d) MCF CONTROLS,A12
cat
dog
horse
cow
sheep
giraffe
leopard
elephant
lion
tiger
monkey
mouse

Figure 2

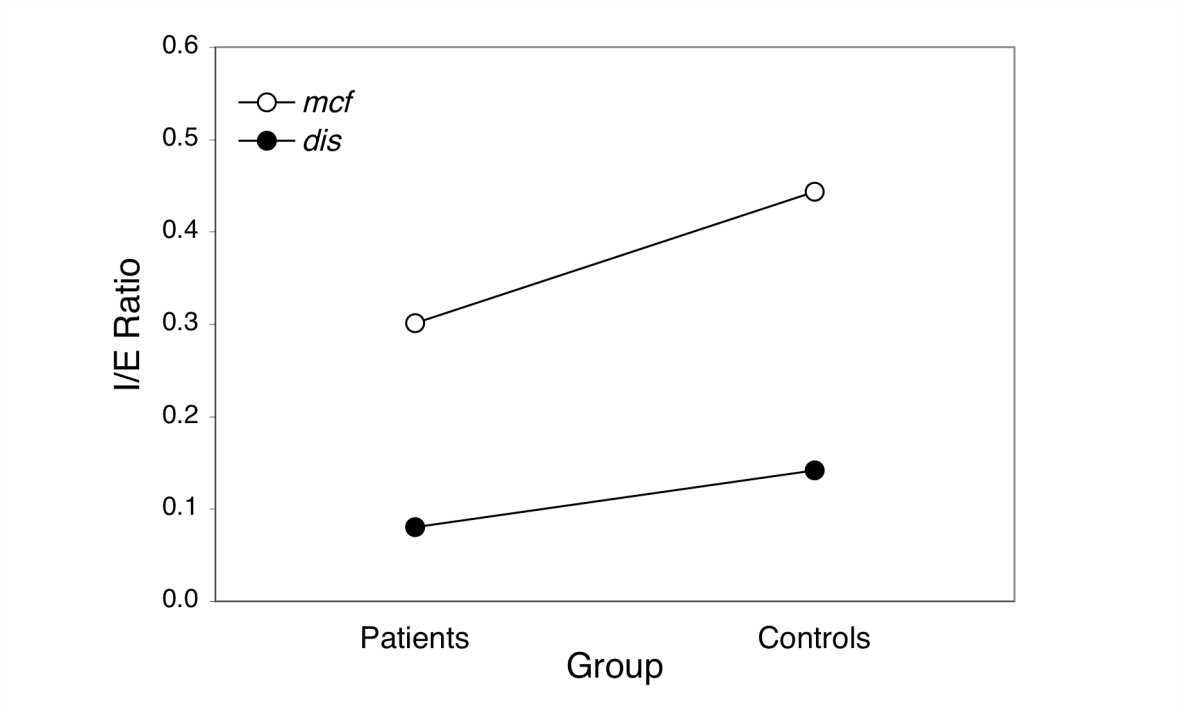


Figure 3

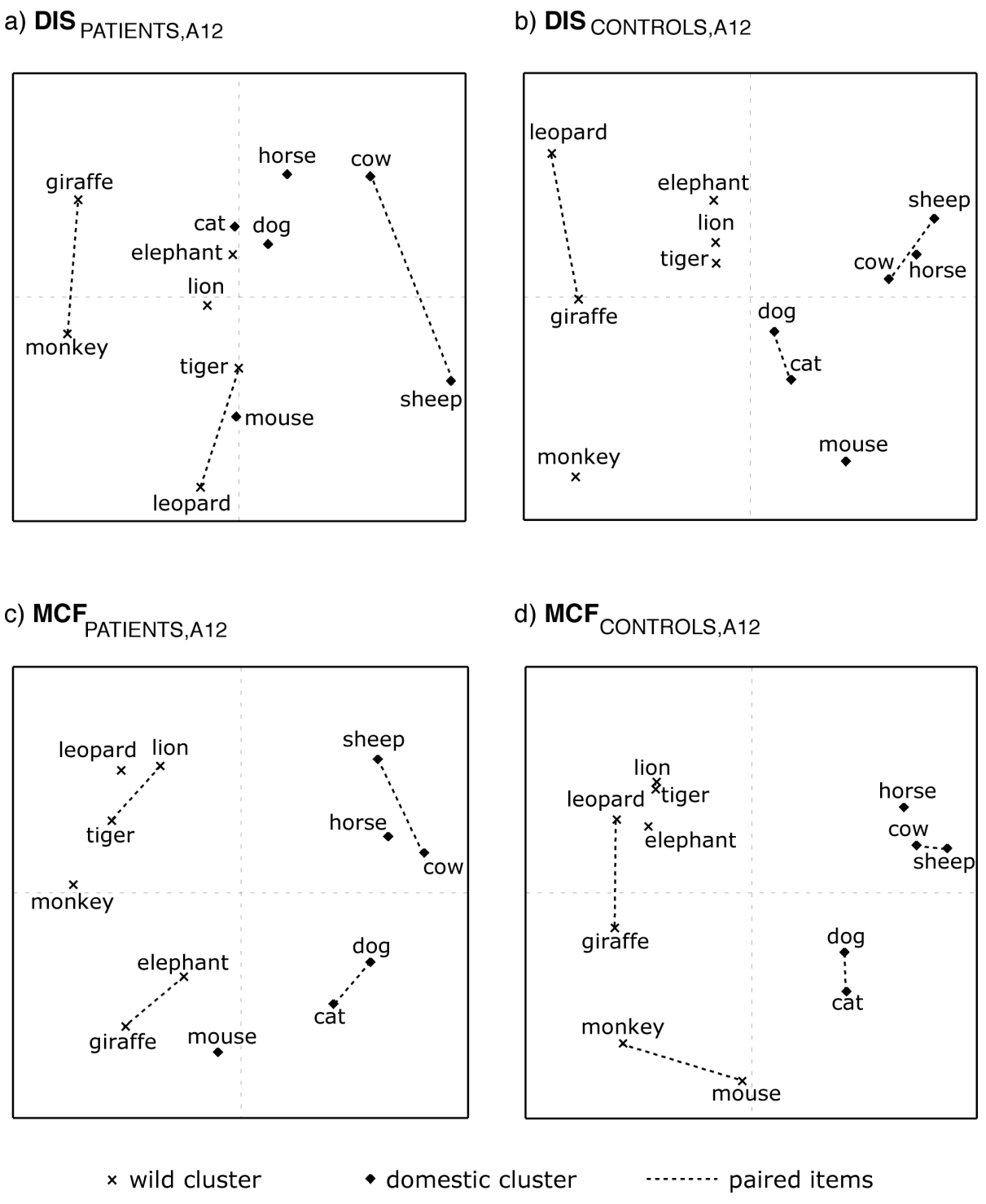

a) DIS PATIENTS,A12
horse
cow
giraffe
cat
dog
elephant
lion
monkey
tiger
sheep
mouse
leopard
b) DIS CONTROLS,A12
leopard
elephant
sheep
lion
tiger
cow
horse
giraffe
dog
cat
mouse
monkey
c) MCF PATIENTS,A12
leopard
lion
sheep
tiger
horse
cow
monkey
elephant
dog
cat
giraffe
mouse
d) MCF CONTROLS,A12
lion
leopard
tiger
horse
elephant
cow
sheep
giraffe
dog
cat
monkey
mouse
× wild cluster
◆ domestic cluster
paired items

Figure 4

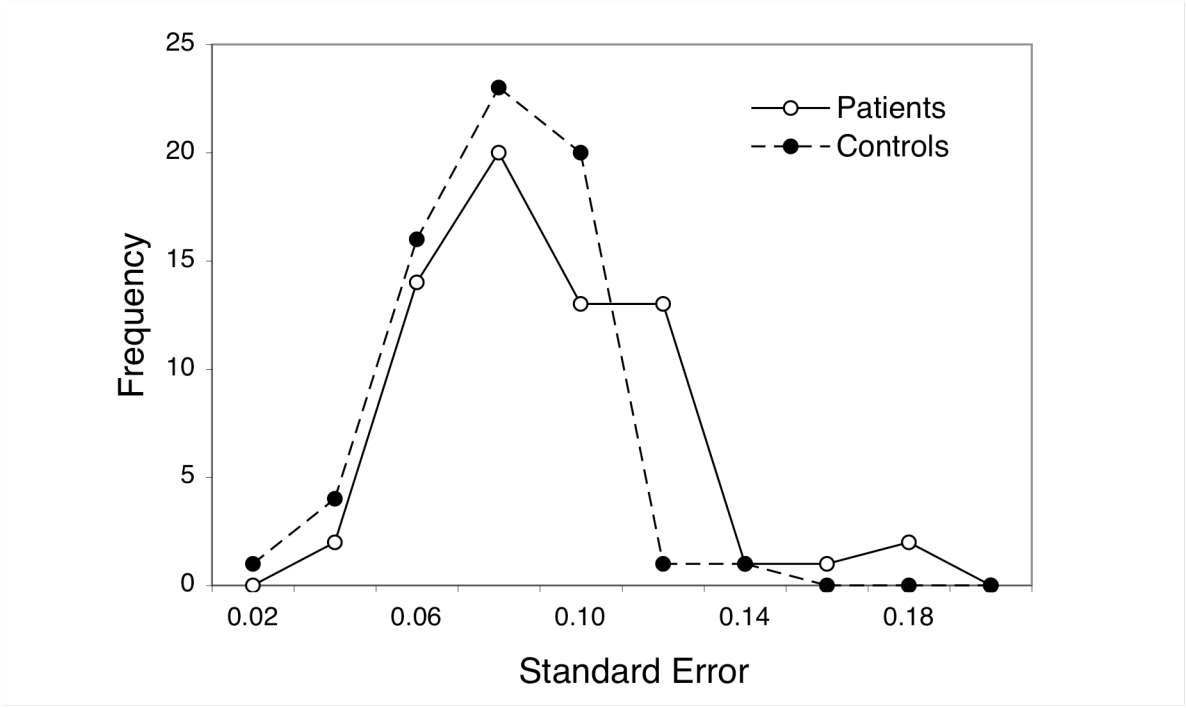

25
20
15
10
5
0
Frequency
0.02
0.06
0.10
0.14
0.18
Standard Error
Patients
Controls

Figure 5

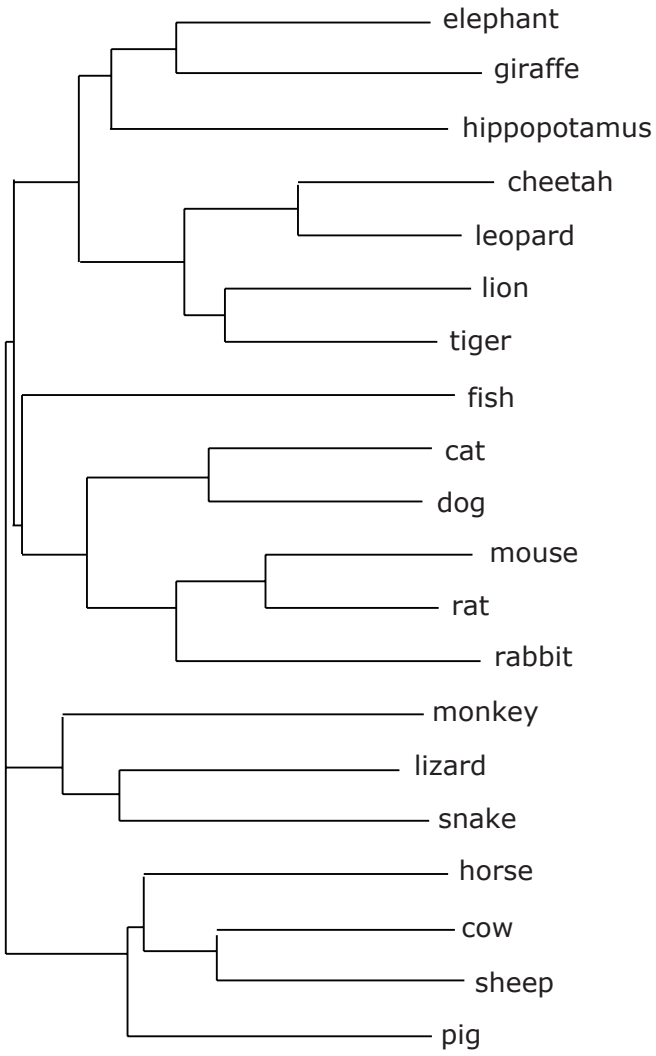
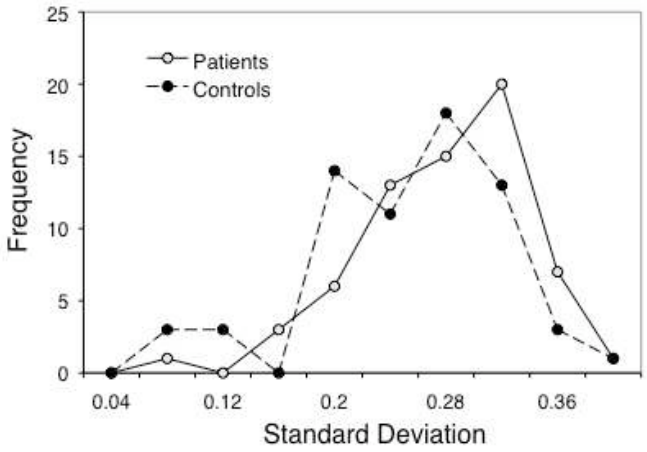

elephant
giraffe
hippopotamus
cheetah
leopard
lion
tiger
fish
cat
dog
mouse
rat
rabbit
monkey
lizard
snake
horse
cow
sheep
pig

Figure 6

Correction to variability analysis (added 07.06.2009)

The first part of our variability analysis (p22 of this document) looked at the standard error of the mean cumulative frequency (*mcf*) scores for 66 item pairs. With hindsight[1] the standard error is not the most appropriate measure for comparing variability between groups since it scales inversely with the square root of the sample size. We now recommend the use of the standard deviation for such comparisons. Reanalyzing our data using the standard deviation shows that, consistent with our original conclusion, variability is somewhat greater for patients than for controls (patient mean=0.259, control= 0.233, F= 5.4, p= .022). The figure below reproduces figure 5 but showing frequency distributions for the standard deviation. In our study, a similar number of item pairs were produced by both groups (802 for patients, 809 for controls) thus the use of the standard error did not unduly affect the between-group variability test, however, similar analyses using unequal sample sizes should not use the standard error. Note that the conclusion of greater variability in the patient data was also supported by split-half comparisons using within- and between-group correlations.



Frequency distributions, over sixty-six item pairs, of the *standard deviation* of the *mcf(G,a,b)* measure for patient and control groups. The graph shows the distance estimates obtained from the patient group are more variable than those of controls.

[1] We are grateful to Daniel Pratt for drawing our attention to this issue.