Deriving semantic structure from category fluency: clustering techniques and their pitfalls

Wouter Voorspoels[1], Gert Storms[1]*, Julia Longenecker[2], Steven Verheyen[1], Daniel R. Weinberger[2], Brita Elvevåg[3,4]

1. Department of Psychology, University of Leuven, Belgium.
2. Clinical Brain Disorders Branch, National Institute of Mental Health/NIH, Bethesda, MD 20892, USA.
3. Psychiatry Research Group, Department of Clinical Medicine, University of Tromsø, Norway.
4. Norwegian Centre for Integrated Care and Telemedicine (NST), University Hospital of North Norway, Tromsø, Norway.

*Corresponding author: Gert Storms Email: Gert.Storms@ppw.kuleuven.be

Abstract

Assessing verbal output in category fluency tasks provides a sensitive indicator of cortical dysfunction. The most common metrics are the overall number of words produced and the number of errors. Two main observations have been made about the structure of the output, first that there is a temporal component to it with words being generated in spurts, and second that the clustering pattern may reflect a search for meanings such that the 'clustering' is attributable to the activation of a specific semantic field in memory. A number of sophisticated approaches to examining the structure of this clustering have been developed, and a core theme is that the similarity relations between category members will reveal the mental semantic structure of the category underlying an individual's responses, which can then be visualized by a number of algorithms, such as MDS, hierarchical clustering, ADDTREE, ADCLUS or SVD. Such approaches have been applied to a variety of neurological and psychiatric populations, and the general conclusion has been that the clinical condition systematically distorts the semantic structure in the patients, as compared to the healthy controls. In the present paper we explore this approach to understanding semantic structure using category fluency data. On the basis of a large pool of patients with schizophrenia (n=204) and healthy control participants (n=204), we find that the methods are problematic and unreliable to the extent that it is not possible to conclude that any putative difference reflects a systematic difference between the semantic representations in patients and controls. Moreover, taking into account the unreliability of the methods, we find that the most probable conclusion to be made is that no difference in underlying semantic representation exists. The consequences of these findings to understanding semantic structure, and the use of category fluency data, in cortical dysfunction are discussed.

1. Introduction

Assessing verbal fluency has a long history within neuropsychology and its clinical value as a sensitive indicator of cortical dysfunction seems indisputable. At its simplest level participants are to name as many words belonging to a certain category (e.g., animals) as possible within a specified period such as a minute. Although of a seemingly straightforward nature, the numerous neurocognitive constructs and processes likely involved in word generation tasks made it an attractive probe of overall mental ability even in the early days of psychometric testing (e.g., Thurstone, 1938; Lezak, 1995). Likely because of their simplicity and brevity of administration, as well as their usefulness as indicators of overall general brain dysfunction, fluency tasks (category and letter) are routinely administered to assess function in a very wide range of neuropsycholological conditions, and most commonly the core metrics are the overall number of words produced and the number of errors (that is, non-members generated for a target category).

The focus of the current paper is category fluency tasks. Two main observations have been made about the structure of the output in these tasks: First, there is a temporal component to it and second, the clustering pattern may reflect underlying semantic mechanisms. Concerning the first issue, it has been noted that words are generated in spurts rather than uniformly in time, and this has been variously modeled (as exponential - Bousfield and Sedgewick, 1944; or hyperbolic - Bousfield et al. 1954). Regarding the second issue, the recall process has been speculated to involve a search for meanings rather than individual items and thus it is assumed that the 'clustering' of words reflect the activation of a specific semantic field in memory (Gruenewald and Lockhead, 1980). Such conceptions are rooted firmly in popular ideas of semantic networks (e.g., Collins and Loftus, 1975; Collins and Quillian, 1969) and in the resulting methodologies with which to assay the speed and efficiency of information search and retrieval from these underlying storage systems putatively arranged as a network (e.g., semantic priming methodology). In the case of category fluency data, many approaches to

examining the structure of the clustering have been developed, as well as calculating the location and frequency of switching to a new subcategory (e.g., in the category 'animals', switching from the subcategory 'domestic' to 'farm'; e.g., Elvevåg et al., 2002; Troyer et al., 1997). However, there are numerous inherent confounders in any methodology that requires so much subjective judgment of cluster boundaries, and indeed Bousfield's concern in 1953 is equally relevant today: "In this situation we cannot rely on the experimenters' subjective judgment, and we would prefer not to rely on the subject's introspections" (p. 229; Bousfield, 1953).

Inspired by the observation that people cluster responses in a category fluency task, a number of studies have focused on the derivation of semantic relatedness, and thus semantic structure, between words (Chan et al., 1993; Prescott et al., 2006; Sung et al., 2012). Two techniques, that form the topic of the present research, have been applied. First, a particularly widely adopted technique to derive semantic structure from verbal fluency consists of calculating the proximity between words during recall (Chan et al., 1993; Prescott et al., 2006). The key intuition underlying this technique is that people cluster similar exemplars of the category in their response order, and thus that the proximity between two items in a response sequence reflects the extent to which these two items are deemed similar. If many items separate the items one is interested in, these items presumably are unrelated and thus not very similar. If few items separate them, the target items are probably rather similar. The similarity relations between category exemplars in turn reveal the mental semantic structure of the category underlying one's responses, which can be visualized by a number of algorithms, such as MDS (Borg and Groenen, 2005; Kruskal and Wish, 1981), hierarchical clustering (Johnson, 1967), ADDTREE (Sattath and Tversky, 1977) or ADCLUS (Arabie and Carroll, 1980). In what follows, we refer to this technique as *VF-PROX*[1].

[1] VF-PROX refers to the use of inter-item proximities (PROX) in a verbal fluency (VF) response sequence to arrive at pairwise similarity.

More recently, a second approach has been proposed, relying on singular value decomposition. Instead of deriving similarity on the basis of inter-item distance in a participant's response sequence, singular value decomposition only takes into account mere co-occurrence patterns of items across participants' response sequences (Sung et al., 2012). That is, if two items often co-occur in response sequences, the analyses will yield a high similarity score for these items, irrespective of their relative position in the sequences. If two items only rarely occur together in the same response sequence, this will result in a low similarity score. Moreover, singular value decomposition would also capture the relatedness between two words that never co-occur together in response sequences, but across sequences do co-occur often with the same words. In the present paper, we refer to this technique as *VF-SVD*[2]. VF-SVD is attributed a number of advantages, in particular regarding the number of items that can be included in the analysis and the dimensionality of the derived representation (we return to this in more detail). Note that VF-SVD is different from more traditional applications of singular value decomposition to derive high dimensional spaces from co-occurrence of words in large text corpora (e.g., Landuaer and Dumais, 1997; for application in the context of schizophrenia, see, e.g., Elvevåg et al., 2007). Indeed, VF-SVD aims at deriving semantic spaces from a relatively small set of word co-occurrence data from response sequences.

Probably due to the ease of administration and availability of category fluency data, the technique of deriving semantic structure from the data has been widely applied in comparisons of semantic structure of patients with various neuropsychological conditions – including Alzheimer's disease and schizophrenia – and healthy control participants (e.g., Aloia et al., 1996; Chan et al., 1993; Chang et al., 2011; Iakimova et al., 2012; Moelter et al., 2001, 2005; Paulsen et al., 1996; Prescott et al., 2006; Rossell et al. 1999; Schwartz et al., 2003; Sumiyoshi et al., 2001, 2006a, 2006b; Sung et al., 2012). The general

[2] VF-SVD refers to using singular value decomposition (SVD) to extract similarity from verbal fluency response sequences.

conclusion of this approach is that a number of neuropsychological conditions systematically affect and distort the semantic structure of the patients, as compared to healthy control participants (but see Elvevåg and Storms, 2003; Storms et al., 2003a and 2003b). VF-PROX has also found its way in other disciplines such as developmental psychology (e.g., Crowe and Prescott, 2003) and cross-cultural psychology (e.g., Winkler-Rhoades et al., 2010).

In the present study, we find the conclusions that follow from application of VF-PROX and VF-SVD to be fundamentally flawed. On the basis of analyses on category fluency data from a large pool of patients with schizophrenia and healthy controls, our data suggest that: (i) Both techniques fail at yielding a reliable measure of inter-item similarity. Neither patient groups nor control groups show sufficient within-group consistency to derive a sensible estimate of the population average, and, consistent with this, the replication reliability is low. (ii) Due to unreliability of the inter-item similarity measure, not only in the patients but also in healthy controls, comparisons make no sense, because the conclusion depends too much on the particular sample and on what is essentially noise in the data. (iii) If we take into account that the data are not reliable, our best estimate, by application of classical psychometric theory, is that the patient group does not systematically differ from the group of healthy control participants.

1.1. Outline

In what follows, we will first present the data that were gathered for the present purpose. We will then demonstrate, separately for the VF-PROX and VF-SVD technique, that the conclusions drawn on the basis of applying the techniques to category fluency data – that is, systematic distortion of the semantic structure due to a specific neuropsychological condition – are flawed. For each method, we start with a brief technical overview and then perform a repetition of earlier research using the data presented, followed by analyses aimed at addressing three questions: (i) Are the similarity data extracted from

category fluency reliable? (ii) Can we make group comparisons on the basis of the extracted similarity data? (iii) What conclusions can we draw taking into account unreliability of the extracted similarity data. Finally, the consequences of our findings for neuropsychology are considered in the general discussion.

2. Data

2.1. Participants

All analyses involve data from a set of 204 patients with schizophrenia and 204 healthy volunteers matched for premorbid intelligence as measured by the Wide Range Achievement Test-Reading (WRAT-R; Jastak and Wilkinson, 1984). All participants were recruited as part of the Clinical Brain Disorders (NIMH) Schizophrenia Sibling Study (DR Weinberger, PI) (Egan et al., 2000). Participants were aged between 21-55 years, free of other medical or neurological problems that might affect performance, learning disabilities, and history of alcohol or drug abuse. Patients were diagnosed by clinicians using the Structured Clinical Interview for DSM-IV Axis I and II Disorders (First et al., 1996). Healthy volunteers received full structured clinical interviews to determine they were free of DSM-IV Axis I and II diagnoses. Participants signed informed consent forms approved for the protocol by the NIMH Institutional Review Board. Age, education, and scores from the WRAT-R and WAIS-R (an estimation of current intelligence from a short form of the Wechsler Adult Intelligence Scale-Revised; WAIS-R – Wechsler, 1981; see also Missar et al, 1994) and corresponding p-values from a one-way ANOVA are listed in Table 1.

Table 1

2.2. Materials

Each participant completed the category fluency task for three different categories (animals, fruits, vegetables) as part of a larger neuropsychological battery. For each category, participants had one minute to generate as many exemplars as they could. They were directed to name any sort of animal, whether it is a group such as "fish" or a species variety such as "rainbow trout". Repetitions and intrusions (non-category words) were not counted in the global score (see Table 1 for score). For the present purpose, we only examined "animals" because there is considerable blurring of semantic boundaries between the other two categories, namely fruits and vegetables (e.g., an avocado and tomato are examples of fruits, but they are often generated as exemplars of the vegetable category; see Storms, De Boeck and Ruts, 2000) and consequently the semantic search process can be expected to be somewhat more complex. Furthermore, the vast majority of neuropsychological studies that used category fluency data to study semantic deficits have focused on animals (Chan et al. 1993; Storms et al., 2003a).

The words were transcribed electronically from hand-written psychometric sheets ~~by the original task administrator~~ in the original order so that we could consider the words in addition to their counts. Instances of identical semantic meanings, but different words (cougar, catamount, puma), or variations in plurality (dog, dogs) were changed to the same form. However, subordinate or superordinate terms were considered unique (e.g. fish vs. trout). Controls generated 303 unique animals, for a total of 4294 words; patients generated 283 unique animals for a total of 3107 words (on average, healthy controls generated more words than patients with schizophrenia, $t=-10.25$, $p<.001$).

3. The VF-PROX procedure

For each participant, a category fluency task provides an ordered list of category exemplars, that is, the response sequence. While many parameters that characterize the response sequence can be fruitfully examined, we focus on extracting information regarding semantic structure on the basis of conceptual similarity data. In the VF-PROX procedure, the similarity data are derived from the response sequences of all participants in a group, in the form of a similarity measure between each pair of items in a set. This procedure has become a widely adopted means of examining semantic structure, particularly in clinical groups (e.g., Aloia et al., 1996; Chan et al., 1993; Chang et al., 2011; Iakimova et al., 2012; Jarrold et al., 2000; Moelter et al., 2005; Paulsen et al., 1996; Prescott et al., 2006; Rossell et al. 1999; Schwartz et al., 2003; Sumiyoshi et al., 2001, 2006a, 2006b) but also in other contexts (e.g., Crowe and Prescott, 2003; Winkler-Rhoades et al., 2010). The key idea is that the underlying, high-dimensional semantic structure is compressed to a one-dimensional sequence of words. On the basis of a number of such one-dimensional sequences (one for each participant who performed the category fluency task), it is hoped that one can derive the underlying semantic structure that is assumed common to all patients (Chan et al., 1993; Prescott et al., 2006) on the one hand, and all control participants on the other hand. Comparison of the underlying semantic structure can then lead to conclusions regarding potential distortions.

More precisely, in VF-PROX conceptual similarity is derived from interitem proximities, that is, the number of words separating two items in a participant's response sequence. For example, when a participant has generated the ordered list {*giraffe, zebra, dog*}, for this participant, the exemplars *giraffe* and *dog* are at distance 2 and the exemplars *zebra* and *dog* are at distance 1. The farther two items are separated, the less similar they are assumed to be. Taking into account length of the response sequence and multiple occurrences in the same sequence, the individual participants' distance scores are combined to form a group mean, the mean cumulative frequency (mcf), formally given by:

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

where D_{abl} is the distance value of participant *l* for exemplars *a* and *b* (see Prescott et al., 2006, for the detailed calculations involved in this, including considerations for repeated words), *G* is the group of participants, *a* and *b* are generated exemplars, and *T* is the number of times *a* and *b* are both included in a participant's response sequence. The resulting distances are considered a measure of dissimilarity between each pair of exemplars, and are thought to reflect the underlying conceptual similarities of the population from which the group is a sample. The similarity scores can then be used as input to several algorithms that rely on proximity data, such as MDS, ADDTREE and ADCLUS. Importantly, however, these algorithms are not the object of our concern; they are merely convenient ways of representing similarity data. The most important aspect of VF-PROX lies in the extraction of pairwise similarity from the response sequences based on interitem proximities.

Our evaluation of the VF-PROX procedure is guided by three questions that are crucial to justify any conclusions: (i) are VF-PROX data reliable, (ii) do group comparisons on the basis of VF-PROX data make sense and (iii) what can we conclude from VF-PROX data regarding the issue of distorted semantics.

3.1. Prelude: An application of VF-PROX

In a first analysis, our aim is an application of VF-PROX in a manner similar to earlier research that has examined differences in semantic structure between patients with schizophrenia and healthy volunteers on the basis of similarity data derived from a category fluency task (e.g., Aloia et al. 1996; Paulsen et al., 1996). This research typically relies on fairly small participant groups of patients and controls (e.g., n=20 per group; we will perform similar analyses for larger samples later). The participants perform a

category fluency task, from which the pairwise similarities for a fairly small set of exemplars (e.g., 12) of a category is extracted following the VF-PROX procedure.

For the present analyses, we follow the exact same procedure. In later analyses, we will illustrate that the VF-PROX procedure does not lead to reliable measurements of similarity and by consequence the observation of differences in pairwise similarity between groups does not warrant conclusions regarding systematic, consistent group differences, let alone conclusions regarding semantic deficits. For now, however, our aim is to observe differences in the MDS-representations of patients and controls, in a way similar to earlier research. The large pool of controls and patients allows us to randomly select a smaller sample of controls and patients, in an identical manner to what is done in a typical study: Instead of going out into the world to find 20 volunteers, we randomly select 20 among the 204 we have available.

3.1.1. A note on sampling

For all following analyses – both in the context of the VF-PROX procedure and later the VF-SVD procedure – it is crucial to appreciate that every single time we sample (for instance, 20 participants) from the large participant pools, the result can be thought of as a new study, as if we would go out in the world and do the study again with different participants. There is no essential difference. Thus, if we sample 10 times from both groups, we have data for 100 virtual studies, since each sample of the one group in combination with a sample of the other group constitutes a repetition. And, by extension, we expect that the results we get from these 100 studies show similar patterns. In the end, we want to infer to population parameters, and by sampling we want to attain good estimates of the population parameters. The population parameters are assumed to be stable unobserved values, and the corresponding sample parameters are expected to deviate from these values, but within acceptable boundaries.

3.1.2. Procedure

From the large group of 204 controls and 204 patients, we randomly sample one group of patients and one group of controls, both of size 20. For all participants, we have available the recorded responses on the category fluency task for the category of animals. For both samples, we performed the VF-PROX procedure to extract similarity data. The reference words were the top twelve animals most frequently recalled by both patients and controls: *bear, bird, cat, cow, dog, elephant, fish, giraffe, horse, lion, snake,* and *tiger*.

3.1.3. Results and discussion

For both the patients and the controls, the dissimilarity-matrix was used as input in a non-metric MDS analysis, which produces, for each group, a geometric representation of the similarity relations between the exemplars of the category. In a geometric stimulus representation, the category exemplars are represented by points, and the distance between points reflects the dissimilarity between the corresponding exemplars (Kruskal and Wish, 1981; Borg and Groenen, 1997). While other tools can be used to represent the dissimilarity data (e.g., tree representations, clustering algorithms, path representations), geometric representations are particularly easy to inspect visually in a simple two dimensional plot. We applied a procrustes transformation to make different MDS-solutions optimally similar without altering the relative distances between each pair of items (e.g., Sibson, 1978). The geometric representations for the patients and the controls are presented in Figure 1.

Clearly, there is some similarity between the geometric representation derived from the patients' category fluency data and the controls' data. In particular, the exemplar pairs *cow-horse*, *cat-dog* and *lion-tiger* are in similar relative position to each other. Closer inspection, however, reveals deviations of the patients group as compared to the controls. As an example, the exemplar pair *cow- giraffe*

(connected by a solid line in Figure 1) presents a difference between both groups. In particular, *giraffe* is in the "wild animals" cluster at the bottom of the controls representations, yet is clearly more in the "domesticated animals" cluster for the patients. Other differences between patients and controls can be observed for *elephant* and *fish*.

Figure 1

As in earlier studies (e.g., Aloia et al., 1996; Paulsen et al., 1996; Prescott et al., 2006), we find differences between the geometric representation of animals in the patient group and the control group. It is thus tempting to draw the conclusion that the underlying semantic structure of patients is systematically different from that of healthy participants: In particular, patients seem to think of wild and domesticated animals in a way that is different from healthy participants. Two important and extremely relevant considerations are appropriate before drawing such a far-reaching conclusion. First, it is not difficult to find differences on a certain criterion between any two groups; the challenge is to find out whether a difference reflects a real population difference or is solely due to random variability. For example, walking in New York one can measure the height of 20 people wearing a dark T-shirt and 20 people wearing a light T-shirt, and find a numerical difference in mean height. The question is whether the observed difference is reliable, which is evaluated by taking into account the variability of height in the populations. Obviously we expect that light-colored T-shirt people are neither smaller nor taller than dark-colored T-shirt people, and in this case, the observed difference is due to the variability of height, which leads to differences in means between imperfect estimates of the population mean.

Thus, it is not the case that, just because a difference is observed, that it necessarily is a meaningful difference.

A second consideration concerns the nature of the differences observed. While earlier studies, and our prelude study, have indeed reported differences between the semantic maps of patients with schizophrenia and healthy controls, little systematicity can be found across studies in the type of differences that are found. If a systematic and consistent difference exists between patients and healthy controls, one would expect the same difference to emerge in most studies. To take the analogy of the T-shirts a step further: If one were to repeat the height study a number of times, one would observe a difference between the mean height of dark T-shirt people and light T-shirt people on every repetition: More precisely, one can expect that in 50% of the repetitions the dark-colored T-shirt people are taller and in 50% of the repetitions the light-colored T-shirt people are taller. While each study shows a difference in mean height, it would be absurd to draw the conclusion that the population of dark T-shirt people has a different height than the population of light T-shirt people. Indeed, one would ascribe the observed differences, which are not consistent across samples, to variability in the population.

In what follows, we will show that the similarity measurements provided by VF-PROX are problematically variable across different samples of the same population. Earlier conclusions regarding differences in semantic structure crucially hang on the assumption that VF-PROX yields a stable and precise measurement of the semantic structure of both patients and controls. If the measurement is not sufficiently precise, the location of the exemplars in the MDS-space is not sufficiently certain, and by consequence, the conclusions are not justified.

3.2. Are VF-PROX data reliable? (i)

If the VF-PROX procedure yields a precise and reliable measurement of conceptual similarity, and thus of a meaningful semantic structure, we expect the position of an exemplar of the category to be relatively

invariant across different repetitions of the task with different participants. The assumption that the sample average converges to the population average lies at the heart of the VF-PROX procedure, and as such, different samples are expected to be very similar. Indeed, this assumption underlies all measurements. If this requirement is not met, for whatever reason, this is problematic for any subsequent analysis (e.g., MDS, ADDTREE, ADCLUS), and a population difference cannot be inferred from an observed difference between samples.

To put the precision and reliability of the measurement of semantic structure to the test, we repeat the study a large number of times on the basis of our large participant pools. In each repetition, we apply a procedure identical to the procedure in the prelude study, which results in a MDS-map of the category animals. Every repetition is a study that could have been performed and reported as the prelude study, and we expect similar results. If patients indeed think of wild and domesticated animals in a fundamentally and systematically different way, we expect a – qualitatively and quantitatively – similar finding to emerge in the large majority of repetitions.

3.2.1. Sampling procedure

A total of 100 random samples of size 20 were drawn from the patient group and the control group (100 samples for each group). For each sample, the exact same procedure as in the previous section was applied to arrive at a geometric representation of the same 12 animals. Again, these 100 samples for each group represent 100 separate studies for a particular group, the equivalent of going out into the world and randomly selecting 20 participants, administering the category fluency task, and performing the VF-PROX analysis to extract similarity data for the population that was sampled. Every combination of a patient and a control sample constitutes a repetition of the comparison made in the previous section.

3.2.2. Results and discussion

To evaluate the reliability of the similarity data extracted from verbal fluency, we used the resulting similarity data of each sample as input in a MDS-analysis to examine the extent to which the position of the category exemplars is invariant across repetitions. Figure 2 presents the geometric representation of the sample of patients and the sample of controls in the previous section. Depicted are the positions of the giraffe for each of the 100 repetitions of the experiment. For reasons of illustrative clarity, we focus on only one exemplar, the giraffe, but similar patterns emerge for every item in the set.

Figure 2

It is clear that the position of the giraffe varies greatly across different repetitions, both for the controls and the patients. Redoing the study with different participants apparently does not guarantee the derivation of a geometric representation in which the giraffe has the same location relative to the other animals. Importantly, differences in location were crucial in concluding that the underlying semantic structure is systematically distorted in patients with schizophrenia. Another sample of 20 patients and controls, however, may have lead to an entirely different conclusion regarding the semantic memory in patients. The giraffe can be considered a wild animal, but also a domesticated one in both populations, depending on the particular samples. The similarity data extracted from the category fluency data are not stable, neither for the patient group nor for the control participants.

3.3. Do group comparisons of VF-PROX data make sense? (ii)

The question is how the lack of reliability in the similarity estimates affects the comparison of these data across groups. Given that the position of exemplars is not reliable in either group, comparisons will lead to fundamentally flawed conclusions, as illustrated in Figure 3.

Figure 3

In Figure 3 the position of the giraffe is projected for all 100 control samples (triangles pointing up) and all 100 patient samples (triangles pointing down). It is clear that control giraffes and patient giraffes are largely among each other, and indeed this is why Figure 3 is difficult to read. It is imperative to realize that any pair of triangles, one pointing up and the other pointing down, represents a repetition of the experiment as presented in the prelude study.

Clearly, the VF-PROX procedure can lead to an array of very different conclusions: We can select a pair for which there is a substantial difference in the position of the giraffe between patients and controls, e.g., the pair that provided the data for our prelude study (indicated by the solid circles in Figure 3). For other samples, however, there is no substantial difference in position of the giraffe. For example, to the right of the control giraffe, one can find a triangle pointing down, referring to the position of the giraffe in a patient sample that shows little difference with the control group in the prelude study. Critically, this particular combination of a control and patient sample would not lead to the conclusion drawn in the prelude study. In other combinations of samples, we can observe a difference between groups in the location of the giraffe, yet of a completely different nature, e.g., a pair of samples in which healthy controls view the giraffe as more domesticated and the patients with schizophrenia consider the giraffe a wild animal. Again, while a difference is observed, this does not

support the findings in the prelude study, due to the completely opposite nature of the difference (keeping in mind the analogy with the dark and light T-shirts). Note that the above does not only apply to *giraffe*, but a similar pattern can be observed for every exemplar in the geometric space.

3.3.1. Discussion

Our analyses have revealed an important limitation of the VF-PROX procedure to uncover semantic structure in both patients and controls. By replicating the experiment 100 times for both patients and controls, we have observed problematic variability, not only for the patients, but also for the controls, in the position of the exemplars in the geometric representations. By consequence, observing differences in location of exemplars across the groups is more a matter of chance than anything else: The differences depend crucially on the particular samples rather than on systematic population differences in semantic memory, whether such differences exist or not. Claiming that groups are different requires replicability of the difference, both quantitatively and qualitatively. Whatever the source of the observed problematic variability[3] across samples – whether it is due to heterogeneous populations or an imprecise measuring methodology –, it leads to unreliable results and thus conclusions that are not justified.

3.4. What conclusions *can* we draw from the present data-set? (iii)

Up to now, we have merely shown that the VF-PROX procedure is insensitive to systematic differences when small samples are considered. So, the question that remains is whether patients have a systematically distorted semantic representation. One logical strategy to remedy variability due to small sample size is to increase the sample size. Following the law of large numbers, we expect the estimation

[3] We return to this issue in the General Discussion.

of the population's semantic structure to improve as more participants are tested. Thus, if systematic differences in semantic memory exist between patients with schizophrenia and healthy comparison participants, larger samples should improve the sensitivity to detect these differences.

A second improvement lies in the use of all data, rather than focusing on only a few category exemplars (e.g., *giraffe*). Even when the data are more reliable, unlikely observations can still occur due to random error. To counter this issue, we focus on the Pearson's product moment correlation coefficient to quantify the relation between the control data and the patient data, which takes into account all pairwise similarities within the set of animals. A near-perfect correlation coefficient indicates that there is no difference in pairwise similarity between the category exemplars, and thus, that there is no difference in semantic representation between the patient and the control participants. The observation of correlations lower than 1 would suggest that differences exist, at least to some extent, again under assumption that the data are reliable.

3.4.1. Increasing sample size

On the basis of our large pool of 204 patients and 204 controls, we can simulate a large number of repetitions, sampling from these pools. For each repetition, we can evaluate the correlation between similarity derived from category fluency data of a patient group and a control group. In general, research using category fluency to extract similarity data relies on fairly small samples of participants. In the present analysis, we will illustrate the effect of increasing the size of the samples drawn from the pool of participants. Figure 4 presents the correlation between control and patient data of a number of repetitions of the experiment, using different sample sizes.

Figure 4

It can be seen that, depending on the particular sample that is drawn, substantially different correlations are obtained, even with sample sizes as large as 100. In one study with, for example, sample size 100, one can observe a correlation of .2 and in another, identical study with different participants, one can observe a correlation of .9. This reflects our earlier finding that VF-PROX may not be ideally suited to extract similarity data from category fluency.

Interestingly, however, as sample size increases, the correlation between control and patient data increases. If we average across all correlations with a given sample size, we find an average correlation of .30 between controls and patients with sample size 20, a correlation of .45 with sample size 50 and .62 with sample size 100. Clearly, even with sample size 100 the correlation suggests that there still is a considerable difference between controls and patients. The general tendency, however, is that the correlation rises as sample size increases. More precisely, the correlations converge to the correlation between the full samples of 204 patients and 204 controls. At the very least, this suggests that patients and controls are more similar in their semantic representation than one might observe on the basis of samples of only 20 people.

3.4.2. Taking into account unreliability

The ultimate question then is whether we find differences in semantic representation as sample size is increased even more. In other words, will there still be differences between the patient and the control group when the data become increasingly reliable?

Relying on the complete pool of 204 controls and 204 patients, the correlation between the two groups amounts to .82, which supports the pattern observed earlier that increasing sample size, and

thus, increasing the reliability of the data, raises the correlation between the groups. But even with as many as 204 participants per group, there still is a difference (i.e., .82 is still different from 1). Given the observed tendency that increasing sample size produces higher correlations, one can hypothesize that adding even more participants would raise the correlation even further, perhaps even arriving at a perfect correlation, implying that no differences exist between the two populations. Indeed, even with a sample size as large as 204, the data are still are not perfectly reliable: The estimated reliabilities, calculated by correcting the split-half correlation with the Spearman-Brown formula (Lord and Novick, 1968), of the similarity data extracted from the verbal fluency task for the controls and patients are .78 and .73, respectively

So, what would be the correlation if we had an infinitely large sample of patient data and an equally large sample of healthy control participants? Phrased differently, what would be the correlation if the data of both groups were perfectly reliable? This can be further examined using classical psychometric techniques (Lord and Novick, 1968). It has been shown that unreliability in variables tends to lower the correlation between two variables. This makes sense, since unreliability is essentially adding random noise, which by definition correlates with nothing. On the basis of this finding, formulas have been developed that allow estimating the correlation under assumption of perfectly reliable data. The formula in question relies on the observed correlation, based on the imperfect data, and the extent to which the data are imperfect, that is, the estimated reliability of the data[4]. Applying the formula, our

[4] The formula to estimate this correlation is: $\hat{r}_{XY} = \frac{r_{XY}}{\sqrt{r_{XX}\, r_{YY}}}$, where r_{XX} and r_{YY} refer to the reliability of respectively X and Y (Lord & Novick, 1968).

best estimate of the correlation amounts to 1[5]. Thus, considering that our data are imperfect, and that this tends to lower correlations, we cannot refute the hypothesis that the correlation is 1.

3.5. Conclusions

Our analyses of the VF-PROX procedure lead to two important conclusions. First, extracting similarity data from a category fluency reflects considerable instability, even when testing up to a tenfold of the number of participants generally inscribed in similar research, and this is the case not only for patients with schizophrenia, but also for healthy controls. Apparently, the VF-PROX procedure unlocks too little systematic information to measure similarity relations within a category with satisfactory precision, and the resulting instability is detrimental for any group comparison: Conclusions on the basis of comparing samples from different populations will generally rely on characteristics of the particular sample rather than on population differences[6].

Second, and contrary to the general conclusion that follows research applying the VF-PROX procedure, the best bet we can make on the basis of the data is that there is *no difference* in semantic representation between controls and patients for the category of animals. This is not to say that we have solid evidence that no differences exist (we return to this in the General Discussion), but it does mean that applying VF-PROX to verbal fluency data does not provide sufficient information to make the claim that there are systematic differences in semantic memory of both groups.

[5] Actually, applying the formula yields a value slightly above 1, due to inevitable unreliability in the estimation procedure.

[6] Moreover, since VF-PROX does not automatically provide information regarding the within-sample variability, the lack of stability cannot be read from its output.

4. The VF-SVD procedure

Recently, a different technique has been applied to verbal fluency data, aimed at answering the same question, that is, whether disorders affecting cortical function lead to systematic distortion of the semantic structure in patients. As in VF-PROX, the aim is to derive a measure of conceptual similarity between category exemplars on the basis of verbal fluency data. Yet, instead of deriving similarity from inter-item proximities, VF-SVD makes use of singular value decomposition. Note that using SVD in this way is crucially different from applications that take large corpora as input for the analysis to derive a high-dimensional semantic space (e.g., Latent Semantic Analysis, Landauer and Dumais, 1997). LSA spaces have already been validated by relating them to behavioral measures of people's performance on a variety of semantic tasks. VF-SVD, on the other hand, used a relatively small data set as input for the singular value decomposition and has not yet been validated. Before presenting a thorough evaluation of VF-SVD, we first provide necessary details on the technique, as applied by Sung et al. (2012).

4.1. Applying singular value decomposition to verbal fluency data

The assumption behind the VF-SVD technique is that if two exemplars are generated by the same participant in a category fluency task, they are similar in one way or another. One can hypothesize that the degree of similarity between two words determines the proportion of participants that will generate the two exemplars in their response sequence. By consequence, if many participants generate the two exemplars, it can be expected that these exemplars have more in common than two items that are only rarely generated in the same response sequence. Put differently, words can be expected to be highly

similar when they co-occur often across response sequences, and highly dissimilar when they rarely co-occur in the response sequences.

In singular value decomposition, the underlying factor structure determining the similarity between all generated exemplars is extracted on the basis of co-occurrence across response sequences. More precisely, an input matrix with rows referring to exemplars and columns referring to participants – and entries denoting whether a participant has generated a particular exemplar – is deconstructed to the product of three matrices that approximates the input matrix, one of the matrices representing the participants in terms of the extracted factors, one matrix representing the exemplars in terms of the factors, and one matrix that links these two matrices. If the number of extracted factors is smaller than the number of exemplars, the dimensionality of the original input matrix is reduced, which is the purpose of SVD in most applications, as this can eliminate error variability. The exemplar by factor matrix contains an "exemplar vector" for each generated item, containing the values of an exemplar on the factors. A measure of similarity is derived in the form of the cosine of the angle between two exemplar vectors (Landauer and Dumais, 1997). The cosine is 1 if two vectors are identical (that is, if two exemplars have identical values across the factors), and 0 if two vectors are orthogonal (that is, if the two exemplars are generated independently across response sequences.

The VF-SVD procedure differs considerably from the VF-PROX technique in a number of respects. Most notably, the input of the singular value decomposition is a participants by items matrix, not encoding rank order information. In other words, whereas VF-PROX procedure extracts similarity on the basis of the co-occurrence of exemplars in a response sequence and their proximity in that sequence, VF-SVD relies only on the co-occurrence of the exemplars across the response sequences of the different participants.

A notable advantage of SVD is that it allows the inclusion of a greater number of category exemplars to evaluate differences in semantic structure. In the VF-PROX procedure, the number of items is limited

because the similarity estimate for a pair of words gets (even more) unreliable if some participants did not generate one or both of the items. Thus, VF-PROX is limited to items that occur in the majority of response sequences (both for patients and healthy controls), a limitation not (explicitly) shared by VF-SVD. Moreover, whereas in general the VF-PROX output is presented in a low dimensional geometric space, Sung et al. (2012) allow a large number of factors in their application. By using a larger number of factors and larger number of items, VF-SVD is claimed to better capture the semantic structure, and thus be more sensitive to differences between groups.

In the following sections, we apply the VF-SVD procedure to our data set, following Sung et al. (2012). Next, we again focus on the three questions addressed earlier. Previewing our results, we find that the VF-SVD procedure suffers from the same problems as VF-PROX: (i) The similarity scores extracted by means of SVD are unreliable, not only for patients with schizophrenia but also for healthy control participants, (ii) by consequence, group comparisons are implicitly flawed, and (iii) if we take into account the unreliability, the VF-SVD procedure provides no convincing evidence that differences in semantic structure exist between patients and healthy controls. The basic line of reasoning is similar to that of the section on VF-PROX, that is, through repeatedly replicating the method we show the instability of the results.

4.2. Prelude: An application of VF-SVD

Sung et al. (2012) gathered verbal fluency data for the category of 'animals' and 'supermarket items' from 102 patients with schizophrenia and 102 controls and after applying the VF-SVD technique, they compared the vector cosines of the 40 most frequently generated exemplars between patients and controls. As in earlier research using VF-PROX, Sung et al. (2012, p. 571) conclude that "category

exemplars reported by persons with [schizophrenia] form less coherent semantic clusters than exemplars reported by healthy adults." We apply the VF-SVD technique with parameter settings identical to those used by Sung et al., restricting our analyses to the category 'animals'.

4.2.1. Procedure

We randomly sampled one group of patients with schizophrenia and one group of healthy controls, both of size 102, similar to Sung et al. (2012), from our larger pool of patients and controls. Their category fluency responses for the category 'animals' were transformed to item by participant matrices, which served as input to the singular value decomposition. For the analyses we used PROPACK (Larsen, 2004). Following Sung et al. (2012) we set the number of factors at 25 and focus on the 40 most generated exemplars (across patients and controls) and compared the similarity values resulting from the cosine of the angle between each two word vectors.

4.2.2. Results and discussion

The correlation between the pairwise similarity scores of the patient and control groups, across all possible pairs, provides a convenient measure of differences in semantic structure between the groups. The observed correlation was .22, at first sight suggesting that there indeed is a difference between the similarity scores extracted from the patient data and the control data. In turn, it is tempting to conclude that this is due to a systematic distortion of semantic memory of the patient groups. Following the same general scheme as in our evaluation of the VF-PROX procedure, we now examine whether this conclusion is valid, keeping in mind that observing a difference does not necessarily reflect a true difference between populations, as illustrated in our T-shirt example. The difference should be replicable, both qualitatively and quantitatively. In what follows we show that this is not the case when using VF-SVD.

4.3. Does VF-SVD yield reliable data? (i)

To evaluate the stability of similarity scores derived through VF-SVD across repetitions within the same population, we repeatedly divide[7] a group into two subgroups of equal size (the size of the subgroups is 102). Each iteration, we perform the VF-SVD procedure for both subgroups separately and derive the cosine similarity scores between all pairs of exemplar vectors. This results in a set of 780 pairwise similarity scores for each subgroup, which can be correlated. The resulting correlation is a measure of reliability, in that a high correlation suggests stability across repetitions within the same population. If the VF-SVD procedure produces reliable output, we expect high correlations between each two subgroups of the same population. The procedure is repeated 500 times for the patient group and the control group. In the two top panels of Figure 5, the histograms of the 500 correlations are shown, one for each group.

The correlation between two halves of a group is rather low, both for the patients and the controls. On average, the correlation is .20 for the controls and .17 for the patients. This means that the similarity scores derived by means of singular value decomposition are extremely unstable across samples of the same population, and by consequence they are bad estimates of the true population means. To make this point more tangible: Doing the analyses on two randomly selected samples of healthy controls would lead to the conclusion that the populations from which the samples are drawn, have different semantics, although they come from the same population, which is of course absurd[8].

[7] We repeatedly divide the groups in two halves instead of drawing a large number of samples, because sampling 102 participants out of our population of 204 would lead to considerable overlap across samples and thus to a rise in correlation simply due to this overlap.

[8] One can argue that it is far from absurd to assume interindividual differences in semantics within the same population. It is, however, an implicit but crucial assumption of both VF-PROX and VF-SVD that there exists a stable population average.

4.4. Do group comparisons of VF-SVD data make sense? (ii)

One could argue that the reported correlation between the patient groups' similarity scores and the control groups' similarity scores is sufficiently small to conclude that the patient groups' semantics are systematically different from the healthy controls' semantics. Yet, it is important to keep in mind that unreliability in the measures essentially is random noise added to the systematic variability, and random noise is not correlated to anything. Low reliability thus results in lower correlations. The question is whether the observed correlation is sufficiently low to conclude that there are differences.

Figure 5

One convenient way to decide whether meaningful differences between groups exist, is to compare the variability within a group with the variability between groups. In the present context, evidence for meaningful group differences exists if the correlation between two samples of the same population (patients or controls) is sufficiently larger than the correlation between samples of different populations. This would indicate that the differences we observe within a group are smaller than differences between groups, which in turn would suggest that the groups are indeed meaningfully different. In more technical terms, we test whether the variability between groups is sufficiently large in the light of the variability within groups to conclude that the observed difference is meaningful (this is very similar to what a t-test would do in our T-shirt example).

We use a procedure identical to that in the previous section. The difference is that, in addition to calculating correlations only between samples of the same group, now we also compute "cross-correlations", that is, correlations between a sample of the control group and a sample of the patient group. Figure 5 presents a visual comparison of histograms of the resulting correlations, within the control group (upper panel), within the patient group (middle panel) and between control and patient samples (lower panel).

With an average correlation of .20 between a control and a patient group, it is clear that the differences in similarity scores between groups are not larger than the differences within group, as the correlations between groups are not significantly different from the correlations within groups (.20 and .17 for controls and patients respectively). This result indicates that similarity scores derived by applying VF-SVD do not warrant the conclusion that systematic distortions in semantic memory of patients suffering from neurological conditions underlie the differences observed. Obviously, given the unreliability of the similarity data, it is near impossible to observe group differences. In the following section, however, we will show that the best bet is that there are no real differences between patients and controls, and that any observed differences are due to variability in the scores that are compared (keep in mind the T-shirt study).

4.5. What conclusion can we draw on the basis of VF-SVD? (iii)

Finally, we make use of all the data available in our data set to make the group comparison, instead of only subgroups, effectively doubling the sample size of that of Sung et al. (2012). Following the law of large numbers, this should raise the reliability, and potentially allow conclusions regarding group differences.

For the 204 patients and the 204 matched controls, we apply the VF-SVD procedure with settings identical to Sung et al. (2012), that is with 25 factors and the 40 most frequently generated items. The

correlation between similarity scores derived from the patient category fluency data and the scores derived from the control category fluency data is .29, which is only slightly higher than in our prelude study. Thus, by doubling the sample sizes, it appears we have uncovered further evidence that the underlying semantic structure of patients and controls are considerably different. Yet, again, the resulting correlation should be interpreted in light of the reliability of the data.

Figure 6

The reliability of the similarity data extracted through use of the VF-SVD procedure is estimated using split-half correlations, corrected by the Spearman-Brown formula. Reliabilities are estimated at .33 and .29 for the controls and the patients, respectively. These values are the average reliability estimate across 500 different divisions of the groups, and they are very low. Similar to our evaluation of the VF-PROX procedure, we can use classical psychometric methods to estimate the correlation between controls and patients were we to have perfectly reliable data.

Figure 6 presents the empirical distribution of the estimated correlation, taking into account that the reliability estimates, and by consequence the estimate of the correlation, depend on the particular split halves one considers. By dividing repeatedly in different halves, we can construct an empirical distribution of the reliability of each group, and of the expected correlation[9].

On the basis of Figure 6, the thesis that a perfect correlation exists between controls and patients cannot be refuted. The distribution of the correlation between controls and patients, assuming we have

[9] This can be easily seen by considering Figures 5, presenting correlations on which reliability analyses are based.

perfectly reliable data, clearly contains 1 (a perfect correlation, implying no differences): While the average estimate of the correlation is .95, which is not perfect, the 95% confidence interval runs from .79 to 1.24[10].

4.6. Conclusions

While at first sight displaying considerable advantages as compared to the VF-PROX procedure, the VF-SVD does not warrant strong conclusions that the observed group differences are due to systematic differences in underlying semantics. Overall, the similarity scores that are derived from VF-SVD are less reliable than the ones extracted through VF-PROX. Moreover, taking into account the unreliability using psychometric and statistical techniques, we find that the procedure does not provide convincing evidence that group differences in semantics between patients with schizophrenia and healthy controls exist. The reason we consistently observe differences in a particular comparison of two samples (as in our prelude study, and in Sung et al., 2012) is the enormous variability across samples, be it samples from different populations or samples from the same population. The observed differences thus arise from random deviations that are sample dependent rather than systematic population differences, and no valid and reliable inferences to the population level can be made.

In light of the presumed advantages of VF-SVD over VF-PROX mentioned earlier (see also Sung et al., 2012), the finding that VF-SVD produces less reliable similarity scores may seem surprising. Yet, it is interesting to consider that essentially, the VF-SVD discards information in the verbal fluency data that is taken into account by VF-PROX; namely the rank-order of generated items. Indeed, Sung et al. (2012) correctly note that the rank-order information can be misleading: In the sequence {*pig*, *goat, cow, lion, tiger}*, the pair *cow-lion* is awarded higher similarity than *pig-cow*. However, one can expect these

[10] The distribution relies on estimates of the true correlation, and can therefore exceed 1.

effects to disappear to a certain extent across a larger number of participants. That is, while most participants will cluster *pig* and *goat*, only rarely will *lion* and *cow* be this close in a sequence.

In a way, the VF-SVD procedure assumes little in terms of cognitive processes that underlie the category fluency task, except that the items in a response sequence are related (which is trivial, since the nature of the task requires them to be related). As can be seen in our analyses, the co-occurrence information is not sufficient to derive pairwise similarity scores. Taking into account that words generated in close proximity are more likely to be similar, as is done in VF-PROX, apparently can be considered an improvement.

This is not to say that singular value decomposition is not useful, on the contrary, it has been successfully applied in a large array of research domains, even in contexts very similar to the present one. Rather, the problem with the present application of singular value decomposition is the data that are entered into the algorithm. Apparently, occurrence in response sequences in a category fluency task simply does not contain sufficient information to capture the underlying semantic similarity between words. However, when used in other contexts – with different input –, singular value decomposition can be a useful instrument. For example, Elvevåg et al. (2007, experiment 2) examined the response sequences of patients with schizophrenia and healthy controls in a category fluency task, and in particular the relatedness of two successive words, measured through the cosine of two words in a Latent Semantic Analysis (LSA) space. The LSA space was derived from text corpora by means of SVD, using as input a large database of word occurrences in text fragments (close to forty thousand text fragments and almost 100000 unique words), resulting in a 300 dimensional semantic space. Likewise, Roll et al. (2012) apply SVD to an even larger corpus containing near 20 million words and successfully use the resulting LSA space to better understand association to cue words in a complex cortical disorder such as Broca's aphasia. Indeed, LSA spaces constructed with SVD on the basis of large text corpora have

been validated by a number of studies, relating it to human sorting and category judgments, similarity judgments, lexical priming tasks and so on (Dumais, 2005).

5. General discussion

The aim of the present article was to evaluate whether two techniques, VF-PROX and VF-SVD, that are used to extract pairwise conceptual similarity from category fluency data, lead to valid conclusions. A prerequisite of the techniques to warrant any conclusion is that they provide a reliable measurement of pairwise similarity. If the estimates are too noisy, that is, if they reflect too much random deviation from the corresponding true population values, they provide an unstable basis to draw inferences, regardless of the subsequent analysis. This is true for any measure of whatever quantity one aims to measure, and thus is a condition *sine qua non* for any measurement and, by extension, any comparison of measurements.

In our analyses we have shown that both VF-PROX and VF-SVD fail to satisfy the condition of producing reliable measurements, to the extent that group comparisons become highly uncertain. Through repeated sampling from a large group of patients and controls, we have revealed that the pairwise similarity scores extracted from category fluency data by means of VF-PROX or VF-SVD vary greatly across samples of the same population, not only for patients, but also for controls. If a population measurement is reliable, one expects it to be stable across different samples (that is, if the underlying characteristics are sufficiently homogeneous in the population, we come back to this later). Moreover, we have demonstrated that this is detrimental for any comparison of the groups in terms of the similarity scores: Depending on the particular sample one considers, a wide range of conclusions can be drawn. If we observe differences using VF-PROX and VF-SVD, these differences emerge due to unreliability, that is, random deviations in the data.

Importantly, we did not only observe problematic unreliability in the patients, but also, and equally so, in the healthy controls. This finding has far-reaching consequences. While one could argue that patients with schizophrenia are more erratic in their response behavior in a category fluency task, which would restrict our findings to this target group, it is highly discouraging to find the same problematic variability in healthy controls. In effect, the present findings generalize to any comparison which involves a group of healthy controls, and thus all comparisons of patients suffering from cortical disorders with healthy control participants. As such, the importance and impact of our results cannot be underestimated: VF-PROX and VF-SVD simply do not yield a reliable measurement of semantic structure, that is, pairwise similarity, on the basis of reasonably sized samples as large as 204 participants, and this is most likely the case for any population in which the techniques have already been applied (e.g., Aloia, et al., 1996; Chan et al., 1993; Chang et al., 2011; Crave and Prescott, 2003; Iakimova et al., 2012; Jarrold et al., 2000; Rossell et al., 1999; Prescott et al., 2006; Schwartz et al., 2003; Sumiyoshi et al., 2006a, 2006b; Sung et al., 2012; Winkler-Rhoades, 2010).

In sum, while our findings do not exclude the possibility that some cortical disorders lead to systematic semantic distortions, they do unmistakingly imply that VF-PROX and VF-SVD are inappropriate, too unreliable, and not sufficiently sensitive to pick up real differences.

5.1. What about idiosyncratic semantic deficits?

One could argue that patients with schizophrenia do have semantic deficits, yet not consistently the same across patients, that is, that the semantic deviations are of a more idiosyncratic nature. Indeed, idiosyncratic deficits would account for the considerable and problematic variability we observed across different samples of patients with schizophrenia. Three considerations are appropriate here. First, earlier research shows that the variability in similarity data in patients with schizophrenia is not consistent across judgments by the same individual made at different times (Elvevåg and Storms, 2003),

suggesting that the variability does not rely on a stable idiosyncratic semantic distortion. Second, we have clearly shown that samples of healthy control participants also reflect a similar variability across samples. This suggests that the problematic variability is a characteristic of VF-PROX and VF-SVD rather than a characteristic of a particular population. Importantly, due to the instability in healthy control participants, we do not have a gold standard to compare an individual patient's deviances with.

Third, it is imperative to appreciate that by yielding mean similarity scores, VF-PROX and VF-SVD are only useful to detect systematic, consistent differences between populations. Both procedures lead to a population estimate for a target population, that is, basically an average value in the population. Even if the procedures were reliable – which is clearly not the case – such an average only makes sense if one assumes the to-be-estimated value is sufficiently consistent across members of the population. Group estimates are sensible only to the extent that participants are inter-individually consistent (see Storms et al., 2003a, for a more elaborate discussion of this issue in the context of patients with cortical dysfunctions). In more specific terms, even if VF-PROX and VF-SVD were reliable techniques, they would only be appropriate to detect when every single patient with schizophrenia would, for example, consider the giraffe a domesticated rather than a wild animal. If, on the contrary, the patients with schizophrenia are heterogeneous, in that different patients differ in different ways from healthy controls (and from each other), treating them as a homogeneous group with a meaningful population average, does not make sense. In this case, data from every participant should be analyzed separately.

5.2. So, are there differences or not?

In the present study, our aim was rather modest, namely to evaluate whether VF-PROX and VF-SVD satisfy a crucially important condition so as to warrant conclusions concerning systematic differences between groups. However, we have also attempted to take into account the unreliability in the data in order to extrapolate what the result would be if the data were perfectly reliable. More precisely, taking

into account the variability due to random noise in the data, the analyses demonstrate, both for VF-PROX and VF-SVD, that the correlation between similarity scores of patients with schizophrenia and healthy controls is not significantly different from 1. Put differently, our best bet, on the basis of the unreliable techniques, is that no systematic differences exist between patients with schizophrenia and healthy control participants.

While the conclusion that no differences exist, in turn is rather tentative due to the enormous instability in the data, it does converge to findings in earlier studies using different techniques. For example, Elvevåg et al. (2005) showed that patients with schizophrenia do not differ significantly from healthy controls in the content and organization of beliefs regarding animals and food. The patients produced similar exemplars in a member generation task, with similar frequencies. Moreover, patients and controls did not differ in their judgments of the member's typicality, and application of the instantiation model (Heit and Barsalou, 1996) to account for the typicality judgments revealed that the organization of the beliefs in patients with schizophrenia paralleled the organization of the control participants. Consistent with this, it has been shown that verbal fluency data of patients with schizophrenia is qualitatively very similar to data from healthy controls, in that the same ideas (i.e., clusters) are accessed, but that the patients's data deviate on a number of parameters because they are slower and less effective at generating ideas (Elvevåg et al., 2002).

5.3. Implications for category fluency data?

Importantly, our results do not show that category fluency data are useless. On the contrary, there are a number of characteristics of response sequences that can be – and have been – usefully examined and compared between patient groups with disorders affecting cortical function and healthy control participants, such as the number of words generated (e.g., Bokat and Goldberg, 2003; Tröster,et al., 1989), the extent to which clusters are exhausted (e.g., Moelter et al., 2001), number of errors, the

association between two subsequently generated exemplars (Elvevåg et al., 2007), and characteristics of the generated words (e.g., Roll et al., 2012).

As to extracting pairwise similarity and semantic structure from verbal fluency data, this seems to be a more complicated matter. Both techniques discussed in our study clearly fail to do so, due to the instability of the measurements the techniques yield. A question that has remained unanswered throughout the present article is the precise origin of this variability. While our conclusions regarding VF-PROX and VF-SVD remain unaltered whatever the origin of the problematic variability in pairwise similarity scores, it is interesting to consider in more detail potential sources of interindividual differences (and even intra-individual differences, Verheyen et al., in preparation). Category fluency data – that is, the particular sequence of words rather than the extracted similarity scores – has been shown only moderately reliable within individuals and even less so between individuals in terms of overlap between responses by the same participant at different times or different participants (Bellezza, 1984). One potential source of inter-individual variability are differences in semantic storage, as is the general, but erroneous, conclusion on the basis of VF-PROX and VF-SVD. Apart from the semantic storage, however, there are numerous cognitive components involved in category fluency, each of which can lead to inter-individual and intra-individual differences. These components belong to two more general classes: Differences in cognitive processes that operate on the semantic representations (e.g., access disorders; see Joyce et al., 1996), and differences in more general cognitive mechanisms (e.g., attention deficits; see Storms et al., 2003b). For example, one can expect that participants vary in the extent to which they are able to exhaust semantic clusters and the relative ease with which they switch clusters (e.g., Elvevåg et al., 2002; Robert et al., 1998), the strategy that is used to select the next cluster, the attention they attribute to the task, the memory they have for exemplars already mentioned, whether they are inclined to revisit clusters after a while, the ability to keep their mind on the task, what general strategy they use and so on.

Keeping in mind these different aspects of generating a response sequence, one can expect a high degree of variability in response sequences across participants, even under an assumption of identical semantic structure. Consequently, the reconstruction of the underlying semantic structure solely on the basis verbal fluency data presents an enormous challenge, one at which VF-PROX and VF-SVD fail. One reason for the techniques' failure is that they are blind to many of the components at work in a category fluency task, and thus cannot accommodate much of the variability in response sequences. For example, VF-SVD is blind to the observation that participants visit and exhaust semantic clusters. While VF-PROX takes into account clustering by relating inter-item distance to similarity, the technique is somewhat blind to the observation that participants switch between clusters: For example, in the sequence {*cow*, *sheep*, *horse*, *whale*, *dolphin*}, the pairwise similarity value attributed to *horse* and *whale* is identical to the value attributed to *cow* and *sheep*.

One path that may lead to success is to implement the different processes that are involved in generating a response sequence in the analyses that are aimed at reconstructing the semantic structure. The potential of this strategy is supported by the present findings, and deserves some elaboration. A quick comparison of the reliability analyses of VF-PROX and VF-SVD reveal that the latter yields similarity data that are even more unreliable than the VF-PROX data (for 204 healthy controls, the estimated reliabilities were .33 and .78, for VF-SVD and VF-PROX respectively). Keep in mind that VF-SVD relies only on the co-occurrence of items across response sequences of different participants, assuming no further process underlying the generation. VF-PROX on the other hand, does take into account clustering to some extent, by considering the proximity between any two response items. Clearly, although far from perfect, the assumed process enables the algorithm to perform better in terms of reliability of the output similarity data.

More elaborate implementation of the processes underlying a response sequence could lead to additional raising of the output similarity data, and thus allow a precise measurement of semantic

structure on the basis of category fluency data. For example, one may expect that the first exemplars in a response sequence rely more on a clustering approach whereas after a while, participants start searching the semantic space more erratically. In estimating the population pairwise similarity between two exemplars, one could choose to attribute more weight to inter-item distances as they appear earlier in a response sequence. Likewise, one might expect that switching clusters, and searching for a new cluster, takes time. The latency between two items can therefore be informative to infer whether participants have switched clusters between two generated items. Such additional assumptions on the processes that underlie the response sequence, will perhaps allow the extraction of more reliable similarity data from category fluency data.

6. Conclusions

Verbal fluency is a convenient measure for assessing the flow of thought and speech. While it undoubtedly offers a window into cortical functioning, and in particular into semantic storage, it is important to appreciate that participants' responses rely on a variety of cognitive and cortical processes that are not merely of a strictly semantic nature (e.g., Bellezza, 1984). In the present paper, we have evaluated two techniques that aim at deriving a measure of conceptual similarity between category members from category fluency responses, in order to compare the semantic memory of patients with cortical dysfunctions with healthy controls. We have clearly shown, on the basis of an extensive sample of patients with schizophrenia and healthy controls, that these two techniques do not yield reliable measurements, and thus lead to highly uncertain conclusions. Importantly, this was the case for both the patient group and the healthy controls. Given the size of our samples, and the robustness of our findings in patients and in controls, we can conclude that the two techniques are not adequate to make comparisons between any two groups (that is, this conclusion does not only apply to comparisons with patients suffering from schizophrenia), on the basis of reasonably sized samples (as large as 204

participants). We propose that, in order to make a measurement regarding semantic memory from category fluency data, techniques should be refined to incorporate more cognitive components that are known to be influential in a category fluency task. Before applying such refined techniques in comparisons of the semantic structure in patient groups with healthy controls, it is imperative that their output is tested for reliability of the outcome measures both in patients and controls.

7. References

Aloia MS, Fourovitch ML, Weinberger DR, and Goldberg TE. An investigation of semantic space in patients with schizophrenia. *Journal of the International Neuropsychological Society*, 2(4): 267-273, 1996.

Arabie P and Carroll JD. MAPCLUS: A mathematical programming approach to fitting the ADCLUS model. *Psychometrika*, 45(2): 211-235, 1980.

Bellezza FS. Reliability of retrieval from semantic memory: Common categories. *Bulletin of the Psychonomic Society*, 22 (5): 324-326, 1984.

Bokat CE and Goldberg TE. Letter and category fluency in schizophrenia patients: A meta-analysis. *Schizophrenia Research*, 64(1): 73-78, 2003.

Borg I and Groenen P. *Modern multidimensional scaling: Theory and applications*. Springer, 2005.

Bousfield WA. The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, 49(2): 229-240, 1953.

Bousfield WA and Sedgewick HW. (1944). An analysis of sequences of restricted associative responses. *Journal of General Psychology*, 30: 149-165, 1944.

Bousfield WA, Sedgewick HW, and Cohen BH. Certain temporal characteristics of the recall of verbal associates. *American Journal of Psychology,* 67: 111-118, 1954.

Chan AS, Butters N, Paulsen JS, Salmon DP, Swenson MR, and Maloney LT. An assessment of the semantic network in patients with Alzheimer's disease. *Journal of Cognitive Neuroscience,* 5(2): 254-261, 1993.

Chang JS, Choi S, Ha K, Ha TH, Cho HS, Chai JE, Cha B, and Moon E. Differential pettern of semantic memory organization between bipolar I and II disorders. *Progress in Neuro-Psychopharmacology and Biological Psychiatry,* 35: 1053-1058, 2011.

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

Dumais ST. Latent semantic analysis. *Annual Review of Information Science and Technology*, 38: 188-230, 2005.

Egan MF, Goldberg TE, Gscheidle T, Weirich M, Bigelow LB, and Weinberger DR. Relative risk of attention deficits in siblings of patients with schizophrenia. *American Journal of Psychiatry,* 157(8): 1309-1316, 2000.

Elvevåg B, Fisher JE, Gurd, JM, and Goldberg, TE. Semantic clustering in verbal fluency: Schizophrenic patients versus control participants. *Psychological Medicine*, 32: 909-917, 2002.

Elvevåg B, Foltz PW, Weinberger DR, and Goldberg TE. Quantifying incoherence in speech: An automated methodology and novel application to schizophrenia. *Schizophrenia Research*, 93(1-3): 304-316, 2007.

Elvevåg B, Heit E, Storms G, and Goldberg T. Category content and structure in schizophrenia: An evaluation using the instantiation principle. *Neuropsychology,* 19(3): 371-380, 2005.

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

First MB, Spitzer RL, Gibbon M, and Williams JBW. *User's guide for the Structured Clinical Interview for DSM-IV Axis I disorders, Research Version, Non-Patient Edition (SCID-I/NP)*. New York Biometrics Research, New York State Psychiatric Institute, 1996.

Gruenewald PJ and Lockhead GR. The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory,* 6(3): 225-240, 1980.

Heit E, and Barsalou LW. The instantiation principle in natural categories. *Memory*, 4(4): 413-452, 1996.

Iakimova G, Serret S, and Askenazy F. P-1246 functional specificities of semantic memory between early-onset schizophrenia and autism-spectrum disorder: Quantitative and qualitative analyses of the verbal fluency task. *European Psychiatry,* 27: Supplement 1, 2012.

Jarrold C, Hartley SJ, Phillips C, and Baddeley AD. Word fluency in Williams syndrome: Evidence for unusual semantic organisation? *Cognitive Neuropsychiatry*, 5(4): 293-319, 2000.

Jastak S and Wilkinson GS. *WRAT-R: Wide range achievement test administration manual.* Western Psychological Services, Los Angeles, 1984.

Johnson SC. Hierarchical clustering schemes. *Psychometrika,* 32(3): 241-254, 1967.

Joyce EM, Collinson SL, and Crichton P. Verbal fluency in schizophrenia: Relationship with executive function, semantic memory and clinical alogia. *Psychological Medicine,* 26(1)*:* 39-49, 1996.

Kruskal JB and Wish M. *Multidimensional scaling*. Beverly Hills; London: Sage Publications, 1981.

Landauer TK and Dumais ST. A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review,* 104(2): 211-240, 1997.

Larsen RM. *PROPACK for Matlab 1.1,* 2004. Retrieved from http://soi.stanford.edu/rmunk/PROPACK/index.html.

Lezak MD. *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press, 1995.

Lord FM and Novick MR. *Statistical theories of mental test scores*. Reading, MA: Addison-Welsley Publishing Company, 1968.

Missar CD, Gold JM, and Goldberg TE. WAIS-R short forms in chronic schizophrenia. *Schizophrenia Research*, 12**:** 247-250, 1994.

Moelter ST, Hill SK, Hughett P, Gur RC, Gur RE, and Ragland JO. Organization of semantic category exemplars in schizophrenia. *Schizophrenia Research,* 78: 209-217, 2005.

Moelter ST, Hill SK, Ragland DJ, Lunardelli A, Gur RC, Gur RE, and Moberg PJ. Controlled and automatic processing during animal word list generation in schizophrenia. *Neuropsychology*, 15(4): 502-509, 2001.

Paulson JS, Romero R, Davis AV, Heaton RK., and Jeste DV. Impairment of the semantic network in schizophrenia. *Psychiatry Research*, 63(2-3): 109-121, 1996.

Prescott TJ, Newton LD, Mir NU, Woodruff PWR, and Parks RW. A new dissimilarity measure for finding semantic structure in category fluency data with implications for understanding memory organization in schizophrenia. *Neuropsychology,* 20(6): 685, 2006.

Robert PH, Lafont V, Medecin I, Berthet L, Thauby S, Baudu C, and Darcourt G. Clustering and switching strategies in verbal fluency tasks: Comparison between schizophrenics and healthy adults. *Journal of the International Neuropsychological Society,* 4: 539-546, 1998.

Roll M, Mårtensson F, Sikström S, Apt P, Arnling-Bååth R, and Horne M. Atypical associations to abstract words in Broca's aphasia. *Cortex,* 48(8): 1068-1072, 2012.

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

Sattah S and Tversky A. Additive similarity trees*. Psychometrika,* 42(3): 319-345, 1977.

Schwartz S, Baldo J, Graves RE, and Brugger P. Pervasive influence of semantics in letter and category fluency: A multidimensional approach. *Brain and Language,* 87: 400-411, 2003.

Sibson R. Studies in the robustness of multidimensional scaling: Procrustes statistics. *Journal of the Royal Statistical Society, Series B,* 40(2): 234–238, 1978.

Storms G, De Boeck P, and Ruts W. Prototype and exemplar based information in natural language categories. *Journal of Memory and Language,* 42(1): 51-73, 2000.

Storms G, Dirikx T, Saerens J, Verstraeten S, and De Deyn PP. On the use of scaling and clustering in the study of semantic deficits. *Neuropsychology, 17(2):* 289-301, 2003a.

Storms G, Dirikx T, Saerens J, Verstraeten S , and De Deyn PP. On what we cannot learn from proximity data. *Neuropsychology,* 17(2): 323-329, 2003b.

Sumiyoshi C, Matsui M, Sumiyoshi T, Yamashita I, Sumiyoshi S, and Kurachi M. 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, 2001.

Sumiyoshi C, Sumiyoshi T, Roy A, Jayathilake K, and Meltzer HY. Atypical antipsychotic drugs and organization of long-term semantic memory: Multidimensional scaling and clustering analyses of category fluency performance in schizophrenia. *The International Journal of Neuropsychopharmacology,* 9: 677-683, 2006a.

Sumiyoshi T, Sumiyoshi CT, Roy A, Jayathilake K, Meltzer HY, and Kurach M. Atypical antipsychotic drugs and organization of long-term semantic memory: Multidimensional scaling and clustering analyses of category fluency performance in schizophrenia. *Annual Report of the Pharmacopsychiatry Research Foundation,* 37: 165-168, 2006b.

Sung K, Gordon B, Vannorsdall TD, Ledoux K, Pickett EJ, Pearlson GD, and Schretlen DJ. Semantic clustering of category fluency in schizophrenia, examined with singular value decomposition. *Journal of the International Neuropsychological Society*, 18: 565-575, 2012.

Thurstone LL. Primary mental abilities. *Psychometric Monographs*, Vol. 1. University Chicago Press, Chicago, 1938.

Tröster AI, Salmon DP, McCullough D, and Butters NA. comparison of the category fluency deficits associated with Alzheimer's and Huntington's disease. *Brain and Language,* 37 (3): 500-513, 1989.

Troyer AK, Moscovitch M, and Winocur G. Clustering and switching as two components of verbal fluency: Evidence from younger and older healthy adults. *Neuropsychology,* 11(1): 138-146, 1997.

Unsworth N, Spillers GJ, and Brewer GA. (2010). Variation in verbal fluency: A latent variable analysis of clustering, switching, and overall performance. *The Quarterly Journal of Experimental Psychology,* 64(3): 447-466, 2010.

Verheyen S. Intra-individual differences in category fluency, (in preparation).

Wechsler D. *WAIS-R manual.* New York: The Psychological Corporation, 1981.

Weickert TW, Goldberg TE, Gold JM, Bigelow LB, Egan MF and Weinberger DR. Cognitive impairments in patients with schizophrenia displaying preserved and compromised intellect. *Archives of General Psychiatry,* 57: 907-913, 2000.

Winkler-Rhoades N, Medin DL, Waxman S, Woodring J, and Ross NO. Naming animals that come to mind: Effects of culture and experience on category fluency. *Journal of Cognition and Culture,* 10: 205-220, 2010.

Table Legend

Table 1. Basic demographics for the two groups, matched for WRAT-R scores. WRAT-R is used as an estimate of putative pre-morbid intelligence in patients with schizophrenia because there is often reported a substantial drop in intelligence from estimated pre-morbid function (Weickert et al., 2000). Mean values and standard deviations are shown for each variable. The bottom three rows refer to fluency data: Letter Fluency is the number of words generated for the letter F, A, and S in three minutes (one minute per word). 'Category fluency general' refers the number of words generated for the categories, "animals", "fruits", and "vegetables" in three minutes (one minute per category). 'Category fluency animals', refers to the number of words generated for animals in one minute.

	patients		controls		
	average	SD	average	SD	p
General information					
Age, yrs	35.51	9.96	32.63	9.44	0.003
Gender, males (%)	156 (76%)		84 (41%)		<0.001
Education, yrs	13.91	1.94	16.02	1.93	<0.001
WRAT-R	102.66	9.62	104.19	8.86	0.095
WAIS-R	91.86	10.51	105.61	9.41	<0.001
Fluency data					
letter fluency	33.59	10.88	42.64	9.37	<0.001
category fluency general	35.99	9.45	50.16	9.41	<0.001
category fluency animals	15.23	4.41	20.43	5.75	<0.001

Figure 1. Geometric representations of 12 exemplars of the category of animals, derived from 20 patients' and 20 controls' responses on a category fluency task. The crosses indicate the position of a particular animal. For one pair *cow-giraffe*, the corresponding points are connected in both groups (solid line).

Figure 2. Geometric representation of the semantic structure of the animal category for the previously sampled group of 20 controls (left panel) and 20 patients (right panel). The crosses refer to the animals as positioned on the basis of the original sample. The points illustrate the location of *giraffe* for 100 repetitions of the experiment for each group. The encircled cross refers to the location of *giraffe* in the prelude study for the respective groups.

Figure 3. Presentation of the category exemplars, indicated by crosses, according to the geometric representation of the original control group. For each of the 100 control samples and 100 patient samples, the location of *giraffe* is projected in the space (after a procrustes transformation). The control giraffes are represented by the upward triangles, the patient giraffes are represented by the downward triangles. The circles represent the location of *giraffe* in the original sample of 20 patients and 20 controls.

Figure 4. Histogram of all correlations between patient and control data on the basis of 100 samples of varying sample size from each group. For example, the upper graph presents the counts of all possible correlations between any pair of a patient and control sample of size 20 (in total this amounts to 10000 correlations: every sample of patients is combined with every sample of healthy controls, resulting in 100 x 100 correlations).

Figure 5. Comparison of the correlations between similarity scores derived from VF-SVD. The upper panel shows a histogram of 500 correlations between halves of the control participant sample. The middle panel presents a histogram of 500 correlations between halves of the patients sample. The lower panel shows the histogram of all correlations between a control group and a patient group (on the basis of the groups used for the upper and the middle panel).

Figure 6. Distribution of the estimated correlation between similarity scores of 204 healthy controls and 204 patients. The histogram reflects the uncertainty in the estimation of this correlation, resulting from the distribution of the reliability estimates across different split halves. That is, the reliability estimates vary somewhat across different iterations of the split halves method. The solid line represents the hypothesis that there are no differences between groups (i.e., the correlation is 1).

Figure 1.

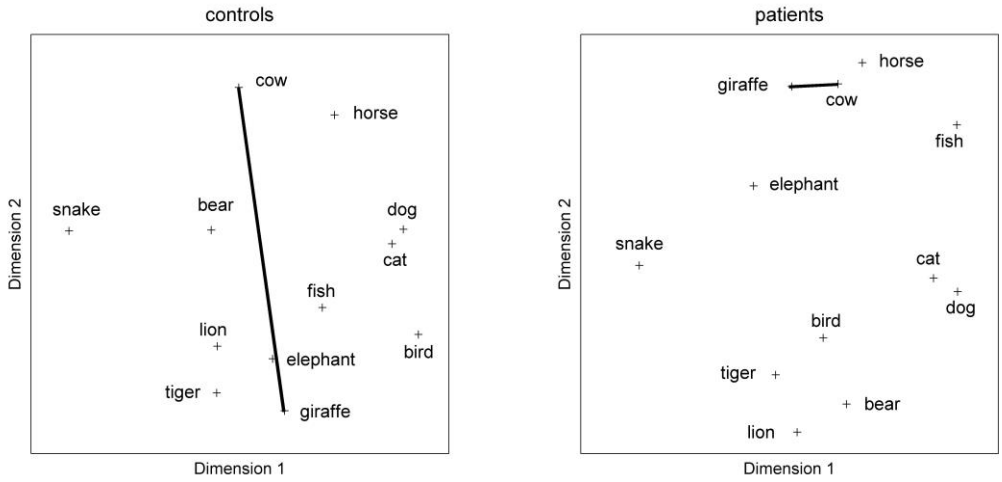

controls
cow
horse
snake
bear
dog
cat
fish
lion
elephant
bird
tiger
giraffe
Dimension 2
Dimension 1
patients
horse
giraffe
cow
fish
elephant
snake
cat
dog
bird
tiger
bear
lion
Dimension 2
Dimension 1

Figure 2.

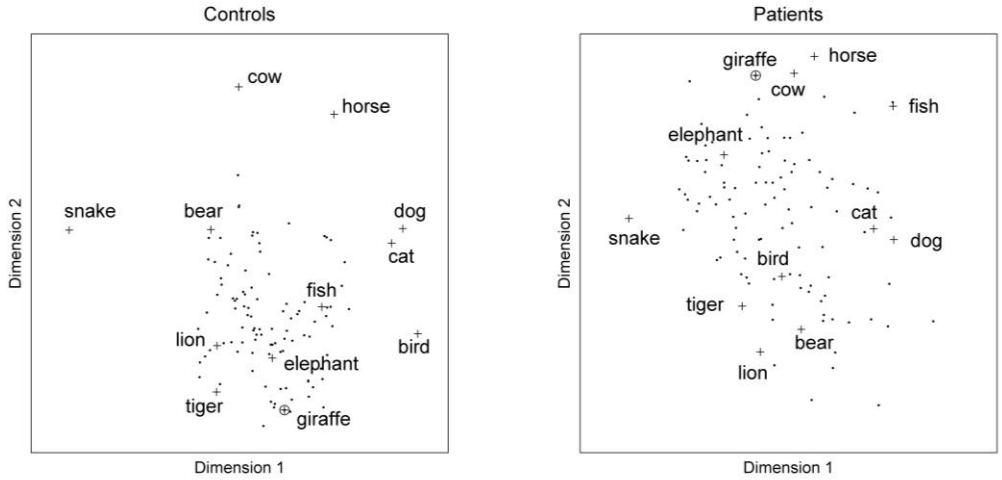

Controls
cow
horse
snake
bear
dog
cat
fish
lion
bird
elephant
tiger
giraffe
Dimension 2
Dimension 1
Patients
giraffe
horse
cow
fish
elephant
cat
snake
dog
bird
tiger
bear
lion
Dimension 2
Dimension 1

Figure 3.

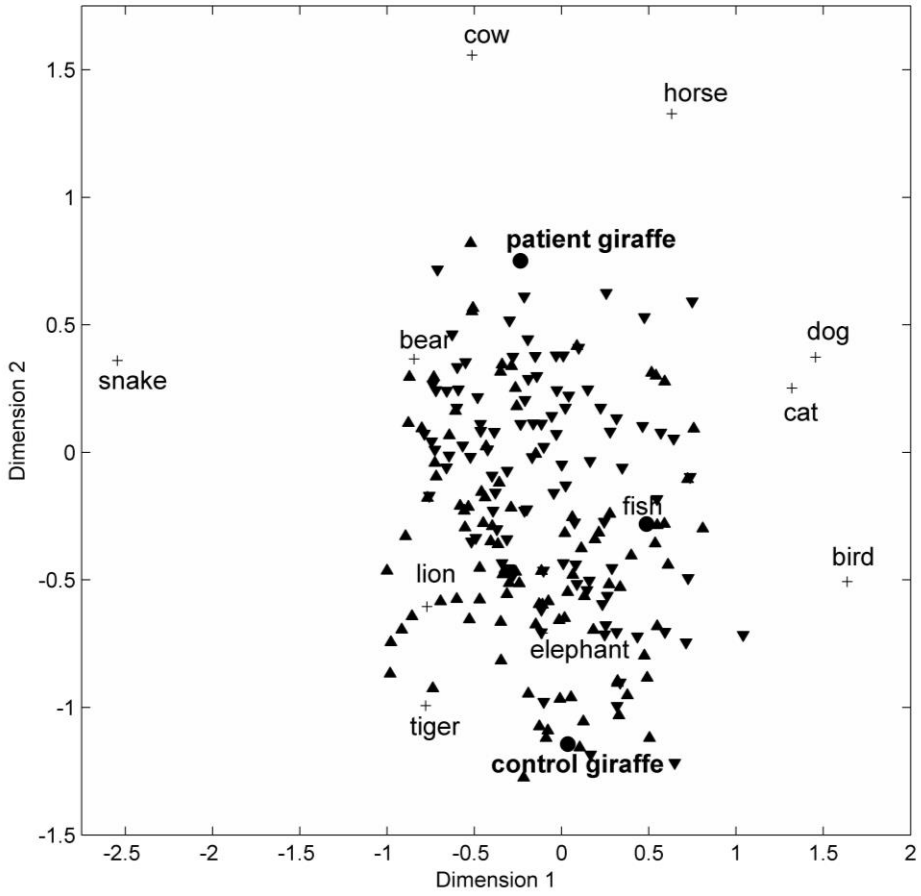

cow
horse
patient giraffe
bear
snake
dog
cat
fish
bird
lion
elephant
tiger
control giraffe
Dimension 2
Dimension 1
1.5
1
0.5
0
-0.5
-1
-1.5
-2.5
-2
-1.5
-1
-0.5
0
0.5
1
1.5
2

Figure 4.

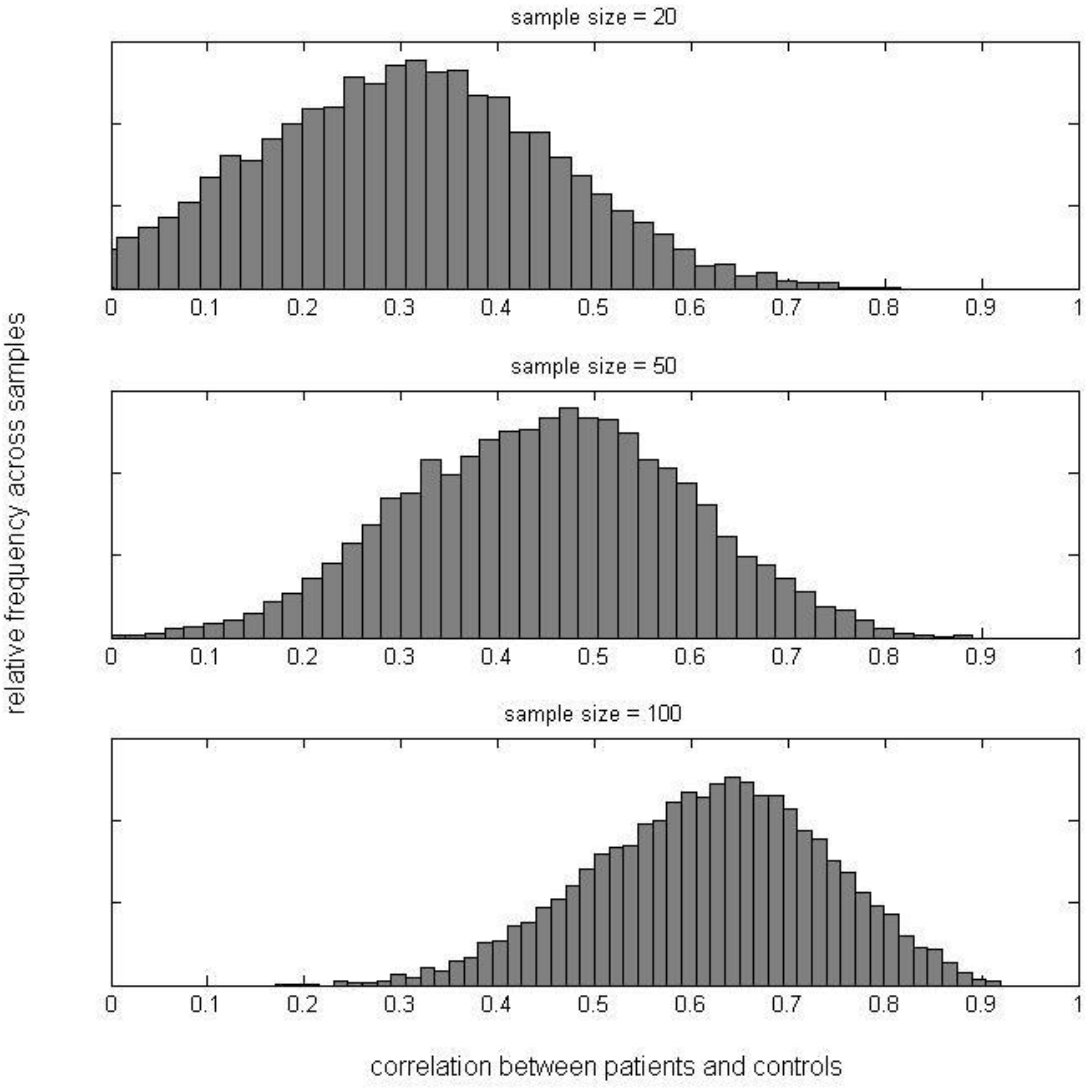

sample size = 20
sample size = 50
sample size = 100
relative frequency across samples
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
correlation between patients and controls

Figure 5.

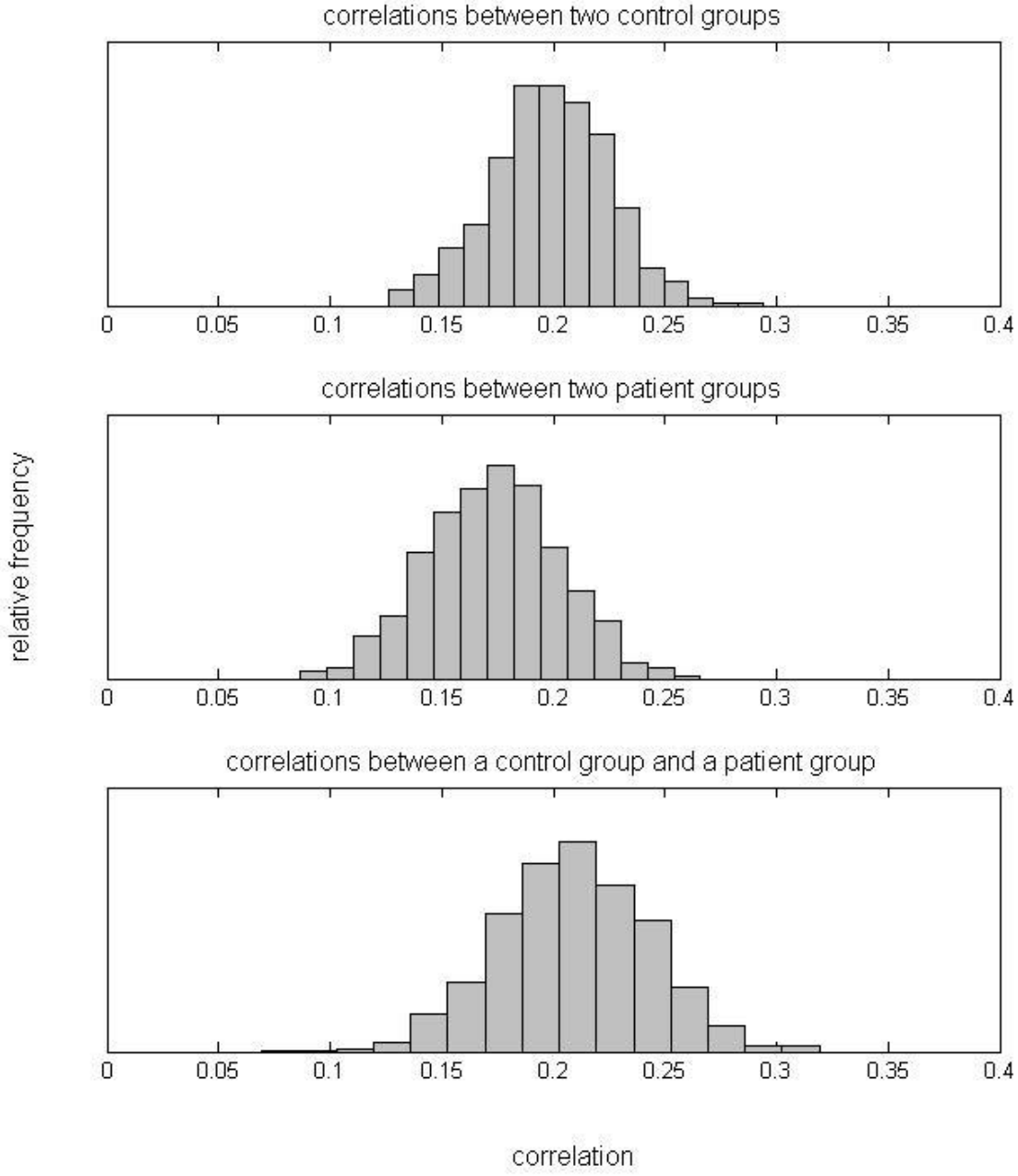
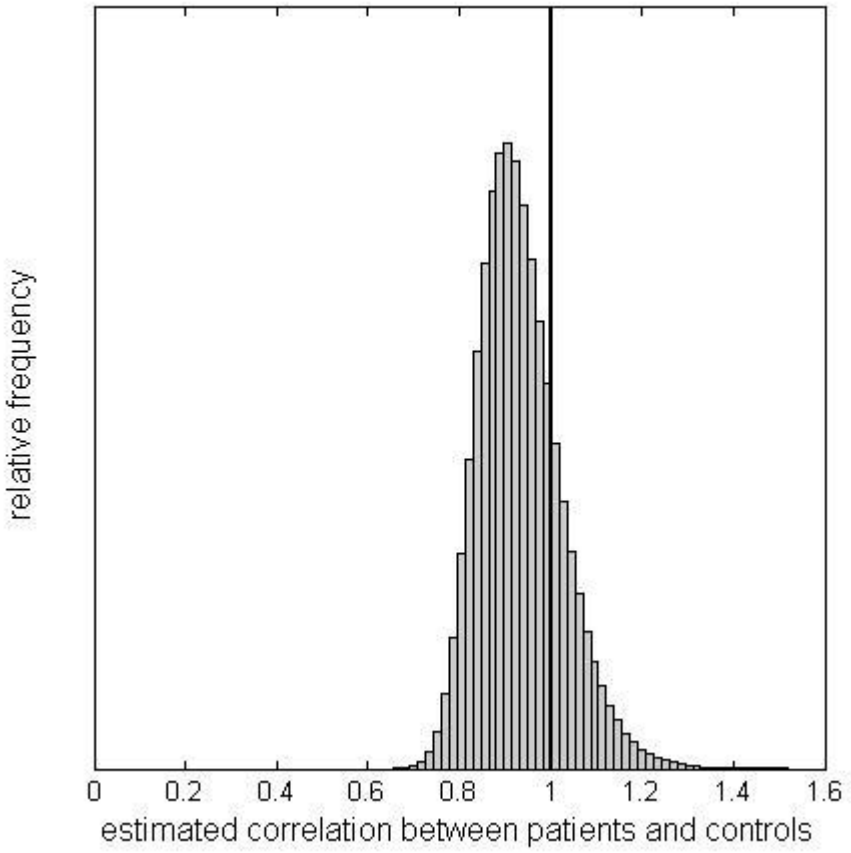

correlations between two control groups
correlations between two patient groups
correlations between a control group and a patient group
relative frequency
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
correlation

Figure 6.


relative frequency
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
estimated correlation between patients and controls