Psychiatry Research 220 (2014) 1125–1130

Contents lists available at ScienceDirect

Psychiatry Research

journal homepage: www.elsevier.com/locate/psychres

Problems of reliability and validity with similarity derived from category fluency

Anne White, Wouter Voorspoels, Gert Storms, Steven Verheyen*

Faculty of Psychology and Educational Sciences, University of Leuven, Leuven, Belgium

ARTICLE INFO

Article history:
Received 16 December 2013
Received in revised form
4 September 2014
Accepted 1 October 2014
Available online 13 October 2014

Keywords:
Categories
Semantic structure
Semantic memory
Spatial representations
Multidimensional scaling

ABSTRACT

This study aims to assess the reliability and the validity of exemplar similarity derived from category fluency tasks. A homogeneous sample of 21 healthy participants completed a category fluency task twice with an interval of one week. They also rated pairs comprised of the most frequently generated exemplars in terms of similarity. Similarities were derived from the fluency data by determining the average distance between generated exemplars and correcting it for repetitions and response sequence length. We calculated the correlation between the similarities derived from the two sessions of the fluency task and between the derived similarities and the directly rated similarities. Spatial representations of the similarities were constructed using multidimensional scaling to visualize the differences between both sessions of the fluency task and the pairwise rating task. We find that the derived similarities are not stable in time and show little correspondence with directly rated similarities. The differences between similarities derived from category fluency tasks in healthy participants, indicate that similar differences between healthy controls and patients with mental disorders, do not necessarily point to a semantic impairment of the latter, but rather reflect the unreliability of the data.

1. Introduction

Similarity is arguably the explanatory construct that is most often invoked to account for the structure of semantic categories like ANIMALS, FRUIT, FURNITURE, and VEHICLES. The similarities between the exemplars of a category are considered to be the proverbial glue that holds a category together. They are what make the category a meaningful and organized whole, rather than a haphazard collection of items. By representing the category exemplars as points in a multidimensional space, whose distances are inversely related to their similarity (through multidimensional scaling or MDS; Borg and Groenen, 2005), the semantic structure of the category becomes manifest (see Verheyen et al., 2007, for an overview). Although direct ratings of the exemplar similarities are usually obtained for this purpose (Dry and Storms, 2009), the belief that the structure of a semantic category can also be reconstructed from category fluency data is wide-held as well. At the heart of this belief lies the assumption that when an individual collapses her multidimensional semantic structure into a one-dimensional sequence of exemplars, she does so by clustering semantically related exemplars: similar exemplars (*cow* and *horse*) are generated closer to each other (within a cluster of farm animals, for instance) than dissimilar exemplars are (*cow* and *lion* across their respective clusters of farm animals and wild animals). According to this line of reasoning, the differences between the ordinal positions of exemplars are adequate measures of the exemplars' similarity (bigger differences indicating smaller similarity) and by subjecting them to a MDS algorithm, the exemplar generation process can be reversed to arrive at the original semantic structure (e.g., Henley, 1969; Chan et al., 1993; Prescott et al., 2006).

The above procedure has often been employed to compare the semantic structures of healthy controls and individuals with mental disorders. A study by Chan et al. (1993) on semantic disruptions in Alzheimer dementia is generally referred to as the prime example of this type of study. Chan et al. asked their participants to generate as many exemplars of the category ANIMALS as possible within a pre-determined time period, computed a measure of exemplar similarity from the fluency lists, and built representations of the category ANIMALS using MDS. The semantic structure of a group of healthy controls was compared with the structure of a group with Alzheimer dementia. Based on several anomalies in the semantic representation of the group with Alzheimer dementia (i.e., individual exemplars that were positioned differently with respect to the other group's representation), Chan et al. concluded that the semantic structure of

* Correspondence to: Faculty of Psychology and Educational Sciences, University of Leuven, Tiensestraat 102, Box 3711, BE-3000 Leuven, Belgium.
Tel.: +32 16 37 30 12; fax: +32 16 32 60 99.
E-mail address: steven.verheyen@ppw.kuleuven.be (S. Verheyen).
URL: http://ppw.kuleuven.be/concat/ (S. Verheyen).

http://dx.doi.org/10.1016/j.psychres.2014.10.001
0165-1781/

patients with Alzheimer diseases is impaired. The procedure has been widely adopted ever since (Aloia et al., 1996; Paulsen et al., 1996; Rossell et al., 1999; Jarrold et al., 2000; Moelter et al., 2001; Sumiyoshi et al., 2001; Prescott et al., 2006; Sumiyoshi et al., 2006, 2009; Chang et al., 2011).

More often than not the application of the procedure to category fluency data from two distinct groups has produced similarities that differ between the groups. However, there is debate about the origin of these differences and the inferences they warrant (Chan and Ho, 2003; Elvevåg and Storms, 2003; Hutchison and Balota, 2003; Jarrold, 2003; Milberg and McGlinchey, 2003; Ober and Shenaut, 2003; Rogers, 2003; Storms et al., 2003a, 2003b; Takane, 2003; Voorspoels et al., 2014). A prime objection to the method relates to a potential lack of reliability of the derived similarities, both for patient and control groups, which might make one erroneously conclude that the semantic structures of two groups differ.

The aim of this study is to evaluate the quality of the similarities derived from the category fluency task, both in terms of their reliability (consistency) and their validity (accuracy). In particular, we assess whether the similarities derived from the fluency data of a healthy group of participants are stable across different measurement occasions and whether they correlate with directly obtained similarities, that is, a gold standard for measuring semantic structure. We asked the same group of volunteers to take the category fluency task twice with a one-week interval. In a healthy homogeneous group a comparison of similarities derived from two identical tasks separated by merely a week should not yield markedly different results: one does not expect the structure of semantic memory to change in a week's time. The choice for a homogeneous group of participants also ensures a fair evaluation of the quality of the measurements, since poor correspondence between measurements then cannot be attributed to random variation among the participants. In addition, if the procedure truly captures a category's semantic structure, one expects high correspondence with the results obtained with an alternative data collection method. Although such a test of the validity of the procedure has been suggested in the past (Chan and Ho, 2003; Ober and Shenaut, 2003) and is also implicitly ascribed to in the literature when different methods are used to obtain similarity data across comparable studies (e.g., Chan et al., 1993 vs. Chan et al., 1995 vs. Ober and Shenaut, 1999), it has not yet been undertaken. Here we used pairwise similarity rating as the alternative task since it is a direct method for obtaining similarity measures that results in better quality data than other methods (Bijmolt and Wedel, 1995; Giordano et al., 2011), is known to render reliable results (Dry and Storms, 2009; Verheyen et al., under review), and allows for the prediction of variables that relate to semantic structure such as typicality, categorization, and induction (Verheyen et al., 2007), which testifies to the method's validity. In addition, 65% of semantic similarity data sets in the literature are obtained through pairwise similarity rating (Dry and Storms, 2009). Taken together, these arguments make the pairwise rating method the gold standard among similarity data collection methods. The quality of the method does come with a price: due to the large number of pairs/comparisons involved, it can be quite taxing and is therefore not generally considered for use among mentally ill patients. Our study allows for the comparison with the pairwise rating procedure, because we rely on healthy volunteers.

For similarities derived from category fluency tasks to be used to study semantic structure, they need to be both reliable (stable in time) and valid (correspond to a generally accepted measure of similarity). The former condition ensures that observed differences can be considered meaningful rather than arbitrary. The latter condition ensures that conclusions pertain to semantic memory. If either condition is unfulfilled, this is a strong contraindication for use of the procedure to study semantic structure in healthy volunteers, but also – as we will argue in Section 4 – for the study of semantic impairments in mentally ill patients.

2. Methods

2.1. Participants

We aimed to obtain a homogeneous sample of participants by recruiting students from the second and third bachelor year of the speech and language therapy program of the University of Leuven. Twenty-one individuals enrolled in the study. All participants were female, aged between 20 and 24 years (mean = 22.11, S.D. = 1.15). A written informed consent was obtained from all participants. They were told that there would be follow-up studies, but they were not informed about the precise content of these follow-up studies. All 21 participants completed the category fluency tasks twice. Nineteen participants also completed a pairwise similarity rating task.

2.2. Procedures

Participants completed a standard category fluency task for four categories: ANIMALS, FRUIT, FURNITURE, and VEHICLES. These four fluency tasks were performed in random order. For each category, participants had one minute to generate as many exemplars as possible. No restrictions were imposed on the exemplars to be generated.

Each participant completed the category fluency tasks on two occasions, with the second session following the first session by a week. During each session data for all four categories were collected from every participant. Identical instructions were used on both occasions.

After six months the participants were requested to perform a pairwise similarity rating task. For ANIMALS, FRUIT, and VEHICLES, the 15 most generated exemplars across both sessions of the fluency task were included in the pairwise rating task. For FURNITURE, 17 exemplars were included because of ties in generation frequency. Participants were asked to rate the similarity of each exemplar pair on a scale ranging from 0 (maximum difference) to 9 (maximum similarity). The categories, the exemplar pairs within a category, and the exemplars within a pair were presented in random order[1].

2.3. Analysis

The fluency outputs were transcribed electronically in the original order. The amount of stemming performed was minimal: plural forms and diminutives were transcribed as one singular form. For each category the 12 most frequent responses across both fluency sessions were selected as targets[2]. For the category ANIMALS the target words were *dog, lion, cat, elephant, tiger, giraffe, monkey, horse, cow, rabbit, fish,* and *crocodile*[3]. For the category FRUIT the target words were *apple, banana, pear, mango, strawberry, tangerine, pineapple, kiwi, grape, melon, orange,* and *lychee*. For the category FURNITURE the target words were *chair, bed, table, closet, couch, desk, office chair, nightstand, wardrobe, coffee table, bench,* and *bookcase*. For the category VEHICLES the target words were *bike, car, bus, plane, train, tram, moped, scooter, truck, boat, metro,* and *helicopter*. Following the procedure described by Prescott et al. (2006) exemplar similarities were derived from the fluency data by determining the average distance between the target exemplars and correcting it for repetitions and response sequence length[4]. This procedure is considered to be superior to earlier proposals by Henley (1969) and Chan et al. (1993). In order to obtain a spatial representation of semantic structure we applied PROC MDS from SAS Version 9.3 to the similarity data using the non-metric, Stress 1, and Euclidean distance options. The results were represented in a two-dimensional space, which is the prevailing practice in the literature (Verheyen et al., 2007). MDS representations of the averaged rated similarities were obtained in the same way. In addition, their reliability was measured using the split-half correlation corrected with the Spearman–Brown formula (Lord and Novick, 1968).

In order to compare the results from both fluency sessions and the results of the pairwise rating task, correlations were calculated between the resulting similarities. An additional comparison was based on visual inspection of the MDS

[1] The decision to assess the method's validity in addition to its reliability was only made after the reliability results were obtained. This took about six months. Thus, the duration of the lag between the fluency tasks and the pairwise similarity rating task is of no particular significance, but merely the result of practicalities involving the organization of the study.

[2] The derivation of similarities from category fluency data requires that each exemplar combination occurs in the response sequence of at least one participant. The largest number of exemplars for which the derivation was technically possible in all four categories was 12. Additional analyses were also performed using 8, 10, and – where possible – 15 or 17 target words. The results of these analyses were similar to the results using 12 target words.

[3] The exemplars *chicken* and *crocodile* have the same response frequency. The results of the dataset containing *chicken* instead of *crocodile* were also analyzed. The results for both datasets were similar.

[4] The procedure described by Prescott et al. (2006) actually yields exemplar dissimilarities varying between 0 (maximum similarity) and 1 (maximum dissimilarity). For ease of presentation these were transformed to similarities by subtracting them from 1. This transformation does not affect any of our analyses.