Psychiatry Research ∎ (∎∎∎∎) ∎∎∎–∎∎∎

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

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/© 2014 Elsevier Ireland Ltd. All rights reserved.

Please cite this article as: White, A., et al., Problems of reliability and validity with similarity derived from category fluency. Psychiatry Research (2014), http://dx.doi.org/10.1016/j.psychres.2014.10.001

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.

Please cite this article as: White, A., et al., Problems of reliability and validity with similarity derived from category fluency. Psychiatry Research (2014), http://dx.doi.org/10.1016/j.psychres.2014.10.001

representations (Fig. 1). In order to facilitate the comparison between the three available representations (fluency session 1, fluency session 2, and the rating task) a Procrustes procedure was applied to the two fluency representations. By means of this procedure the representations of session 1 and session 2 were adjusted to resemble the representation of the rating task in the best possible way by a set of transformations that are admissible for MDS solutions (Borg and Groenen, 2005). The three configurations were compared in terms of the relative positions of the exemplars, the composition of exemplar clusters, and the constituting dimensions.

3. Results

The correlations between the similarities derived from fluency session 1 and from fluency session 2 can be found in the second column of Table 1. These correlations signal that the variance shared by the data from the two sessions is small, ranging from 13% for VEHICLES to 49% for ANIMALS. The third and fourth column

Table 1
Correlations between derived (fluency1, fluency 2) and rated similarities.

	Fluency 1–fluency 2	Fluency 1–rated	Fluency 2–rated
ANIMALS	0.70	0.30	0.34
FRUIT	0.51	0.02	0.00
FURNITURE	0.61	0.03	0.00
VEHICLES	0.37	–0.01	0.16

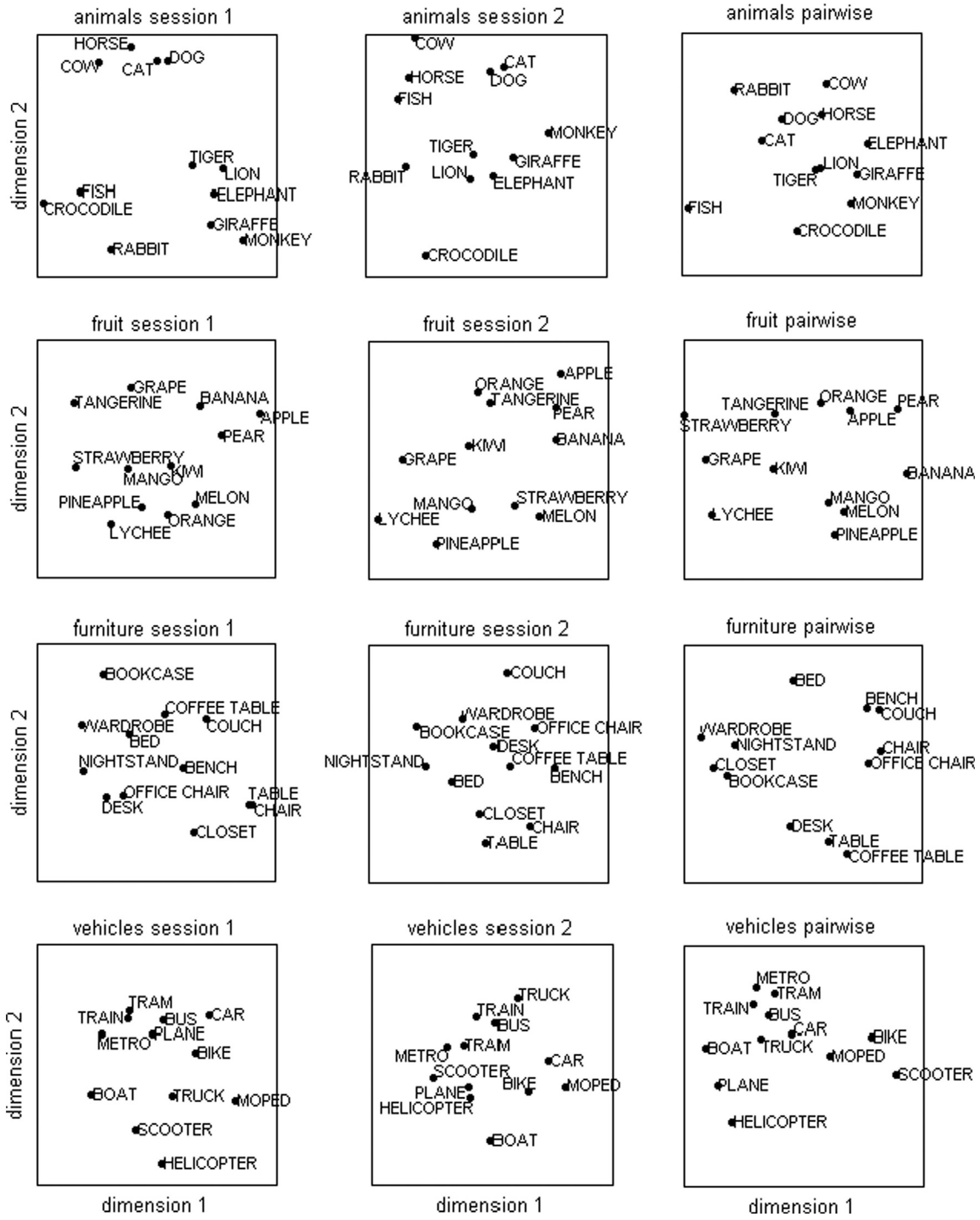

Fig. 1. Two-dimensional MDS representations of the similarities derived from fluency session 1 (left), of the similarities derived from fluency session 2 (middle), and the directly rated similarities (right) for the categories ANIMALS, FRUIT, FURNITURE, and VEHICLES.

Please cite this article as: White, A., et al., Problems of reliability and validity with similarity derived from category fluency. Psychiatry Research (2014), http://dx.doi.org/10.1016/j.psychres.2014.10.001

of Table 1 hold the correlations between the rated similarities and the derived dissimilarities from session 1 and session 2, respectively. They signal that there is hardly any variance shared by the data from the two tasks. All these correlations are significantly different from the reliabilities of the directly rated similarities, which measure 0.96, 0.89, 0.96, and 0.98 for ANIMALS, FRUIT, FURNITURE, and VEHICLES, respectively. These values signify the upper boundary for correlations of external variables with the rated similarities and as such constitute the target values for the correlations between the derived and rated similarities, which are supposedly measuring the same thing (Lord and Novick, 1968)[5].

One can observe marked differences between the MDS representations of the similarity data. The positions of the exemplars are not consistent over the different tasks. For example, in the MDS representation of the rating data *rabbit* is closest to the domestic ANIMALS *cow*, *horse*, *cat*, and *dog* (Fig. 1, upper right panel), while in the MDS representation of the first fluency session it is closest to the aquatic ANIMALS *fish* and *crocodile* (Fig. 1, upper left panel) and in the MDS representation of the second fluency session it is closer to wild ANIMALS like *lion* and *tiger* (Fig. 1, upper middle panel). The *fish* is another example of an exemplar that is positioned differently, both when the two fluency sessions are compared and when the separate fluency sessions are compared with the rating task. It is not clustered with any other exemplar in the representation of the rating task. In the first fluency session, however, *fish* is clustered with *crocodile*, and in the second session with *cow* and *horse*. Similar observations can be made for exemplars from any of the three other categories. The differences are even more dramatic as there is less correspondence for FRUIT, FURNITURE, and VEHICLES than there is for ANIMALS, both between derived similarities and between derived and directly rated similarities (see Table 1).

The differences between the positions of individual items affect the global structure of the MDS representations as well. While the ANIMALS rating data support an organization of the exemplars along the dimensions size (horizontal) and domesticity (vertical), the varying position of items such as *rabbit* and *fish* precludes such an interpretation of the dimensions of the MDS representations of the fluency data. In the representation of fluency session 1, for instance, the domesticated ANIMALS *rabbit* and *cow* are positioned at opposite ends of the vertical dimension. A similar observation holds for MDS representations of categories like FURNITURE, that are organized in terms of clusters rather than dimensions. The organization in terms of closets, seats, and tables of the MDS representation of the rating data, is completely absent in the MDS representations of the fluency sessions (Fig. 1, third row panels). Both examples also illustrate that while the MDS representations of the ratings tend to have face validity in that the organization of the exemplars appears sensible, this tends not to be the case for the MDS representations of the fluency sessions.

4. Discussion

The purpose of this study was to assess whether exemplar similarities derived from category fluency tasks are reliable (consistent) and valid (accurate). To this end the stability in time of the derived similarities was evaluated in a homogeneous group of healthy individuals. If the derived similarities truly capture the structure of a semantic category, two measurements separated by merely one week should yield nearly identical if not very similar results. Moreover, there should be a strong resemblance between similarities derived from category fluency and "gold standard" similarities. To evaluate this, we asked the individuals who participated in both fluency tasks to rate the similarity of pairs of exemplars. The derived similarities should correlate highly with the rated similarities. Fulfilment of both these conditions is a prerequisite for the meaningful use of the similarity derivation procedure, for instance with the intent to compare the semantic structures of a healthy control group and a patient group.

Concerning the stability in time one need not doubt whether the correlations between the similarities derived from the two fluency sessions should be considered sufficiently high or not. Such doubts are immediately resolved when one compares the MDS representations of the resulting data in the same way one would do when comparing data from two distinct groups of individuals. Abnormal associations, changes in relative positions of exemplars, and/or changes in cluster composition would then be taken as arguments for the existence of a semantic impairment in one of the groups (e.g., Chan et al., 1993). All these observations can also be made in the current study with healthy individuals who were evaluated with merely a week in between. The conclusion of a semantic impairment in this case is clearly not feasible, unless one wishes to accept that the semantic structure of a healthy group of participants has been significantly disrupted in the course of one week. A less dramatic explanation of the session difference would be to attribute it to experiences the participants might have had between the two test moments. However, this would constitute a clear concession that exemplar retrieval order in the fluency task is not a pure reflection of semantic similarity, which is considered to be stable rather than ephemeral. Moreover, it would require one to come up with an explanation (in terms of particular experiences) not only of why *fish* and *crocodile* tend to be retrieved further apart in the second fluency session than in the first, but also of the various other changed associations we observed across four different semantic categories. A more parsimonious explanation is that these changes are not systematic, but the result of an unreliable procedure for arriving at similarity data.

The comparison of the similarities derived from the category fluency tasks with directly rated similarities eradicates the possibility that, due to repetition of the task, the session 2 data are somehow negatively affected but the session 1 data are not. Both the similarities derived from session 1 and from session 2 correlate poorly with the rated similarities and both yield MDS representations that differ in important respects from the representation of the rated similarities. That the shortcoming lies with the derived similarities follows from the fact that the rated similarities proved highly reliable, corresponded closely to normative similarities (see footnote 5), and yielded MDS representations with higher face validity than the derived similarities did.

All these elements lead to the conclusion that similarities derived from the category fluency task have problems of reliability and validity. The average distance between exemplars generated in a category fluency task is an unreliable measure in that identical repetition of the measurement yields widely different results. Nor is it a valid measure: the resulting differences in the average distance between two exemplars are inconsistent with the relative differences as assessed by a method that directly captures the actual similarities. The derived similarities can therefore not be relied upon to study semantic structure. What might appear as interesting patterns in the data, might be meaningless fluctuations that have no bearing with semantic memory.

It is important to appreciate that none of these conclusions result from the choice of stimuli. The results were established across four different semantic categories and by choosing the most

---

[5] To ensure that the semantic structures of our participant sample are representative, we correlated their pairwise similarities with normative similarities taken from De Deyne et al. (2008). For the categories FRUIT and VEHICLES – for which there is perfect overlap of the materials – the correlations are 0.84 and 0.96, respectively. These values are very close to the reliabilities of the pairwise similarities provided by our participants. This shows that their semantic structures can be generalized to other participant samples.

Please cite this article as: White, A., et al., Problems of reliability and validity with similarity derived from category fluency. Psychiatry Research (2014), http://dx.doi.org/10.1016/j.psychres.2014.10.001

frequent exemplars we ensured inclusion of the stimuli for which the most observations were available. Neither are the issues resolved by merely increasing the number of participants. Voorspoels et al. (2014) obtained category fluency data for the category of ANIMALS from 204 healthy volunteers. This unusually large number of participants allowed them to conduct a simulation study in which they drew random samples of size N from the 204 participants and compared the derived similarities for each of the samples. Even with samples of 100 participants, considerable differences between the derived similarities were observed. Nor do the issues depend on our particular participant sample of female speech therapy students. While we selected a homogeneous and thus specific group of participants, they produced representative data: their rated similarities correlate strongly with normative similarities provided by male and female university students from a different program (De Deyne et al., 2008; see footnote 5) and their fluency data demonstrate characteristics that are similar to those provided over 30 years ago by students living on another continent, speaking a different language (Bellezza, 1984; see Supplemental material for details). The simulation study by Voorspoels et al. (2014) reached a similar conclusion regarding the reliability of the derived similarities using a more diverse group of participants comprised of both male and female volunteers, aged 21–55 years, with various educational backgrounds.

The issues we identified pose a serious problem for any application of the similarity derivation procedure, but are likely to be even more pronounced when the procedure is applied to patient data, as it tends to be more variable than that of their healthy counterparts (Storms et al., 2003a). Therefore, if a group of patients with idiosyncratic semantic impairments is studied, our contention would be that the problems of reliability and validity would be even more pronounced. It is difficult if not impossible to imagine how increased diversity in a participant sample would increase the reliability of the resulting data. Indeed, the reason why we opted for a homogenous sample in the current study is so that poor correspondence between measurements could not be attributed to random variation among the participants. For a discussion of whether it is sensible to consider deriving similarities for a patient group as a whole in the first place, see Elvevåg and Storms (2003) and Storms et al. (2003a).

The overall conclusion of this study is that the current procedures for deriving similarities from category fluency tasks are not suited to study semantic structure. They suffer from problems of reliability and validity, as evidenced by insufficient stability in time and insufficient correspondence with a gold standard of semantic structure, respectively. This conclusion was arrived at with data from healthy participants, but we believe it applies at least to the same extent to data from patient groups. The implications of this study are clear: the conclusions that in the past have been made on the basis of similarities derived from category fluency are highly uncertain. Reported differences between groups of patients and healthy controls may not be substantial and may have nothing to do with semantic memory. The methodology does not allow decisions concerning the presence of a semantic impairment, let alone conclusions concerning the nature of this impairment.

The very differences in similarities derived from category fluency that in the past have been observed between healthy controls and the mentally ill, can also be observed in healthy participants that are measured twice. These differences should thus not be regarded indicative of a semantic impairment, but rather reflect the unreliability of the used method. Claims pertaining to the existence of semantic impairments in particular patient groups that have been made using similarities derived from category fluency need therefore to be re-evaluated with more sensitive methods. Based on the requirements proposed by Warrington and Shallice (Shallice, 1988), Storms et al. (2003a) already sketched the ideal study aimed at establishing a semantic storage deficit (as opposed to the impaired retrieval of semantic information): The evidence for such a claim can never come from similarity data alone. It needs to be shown that semantic cues do not help patients to access conceptual information and that detailed conceptual knowledge is lost, while superordinate level knowledge is preserved. Only when these two requirements are met and an individual patient's similarity data prove consistent across two measurement occasions, can deviations between the patient's data and data from healthy control participants considered evidence for a semantic impairment. Our findings make it clear that one should not rely on similarities derived from category fluency in such a study. The derived similarities do not capture healthy controls' semantic structure. For use with patients, the pairwise similarity rating task may also be ill-suited because of its tedious nature. When considering alternative procedures for measuring similarity, such as triadic comparisons (Chan et al., 1995; Sylvester and Shimamura, 2002) or spatial arrangement (Ober and Shenaut, 1999; Moelter et al., 2005), it is advised to undertake a reliability and validity analysis of the kind proposed in this paper, to ensure that the normative data to compare patients' similarity data against truly reflect semantic structure.

Acknowledgment

Wouter Voorspoels and Steven Verheyen are both postdoctoral fellows at the Research Foundation – Flanders. We thank Simon De Deyne and Wolf Vanpaemel for helpful suggestions.

Appendix A. Supplementary materials

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.psychres.2014.10.001.

References

Aloia, M.S., Gourovitch, M.L., Weinberger, D.R., Goldberg, T.E., 1996. An investigation of semantic space in patients with schizophrenia. Journal of the International Neuropsychological Society 2, 267–273.

Bellezza, F.S., 1984. Reliability of retrieval from semantic memory: common categories. Bulletin of the Psychonomic Society 22, 324–326.

Borg, I., Groenen, P., 2005. Modern Multidimensional Scaling: Theory and Applications. Springer, New York.

Bijmolt, T.H.A., Wedel, M., 1995. The effects of alternative methods of collecting similarity data for multidimensional scaling. International Journal of Research in Marketing 12, 363–371.

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

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

Chan, A.S., Ho, Y.C., 2003. Things aren't as bad as they seem: a comment on Storms et al. (2003). Neuropsychology 17, 302–305.

Chang, J.S., Choi, S., Ha, K., Ha, T.H., Cho, H.S., Choi, J.E., Cha, B., Moon, E., 2011. Differential pattern of semantic memory organization between bipolar I and II disorders. Progress in Neuro-Psychopharmacology and Biological Psychiatry 35, 1053–1058.

De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M.J., Voorspoels, W., Storms, G., 2008. Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. Behavioral Research Methods 40, 1030–1048.

Dry, M.J., Storms, G., 2009. Similar, but not the same: a comparison of the utility of directly-rated and feature-based similarity measures for generating spatial models of conceptual data. Behavior Research Methods 41, 889–900.

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

Giordano, B.L., Guastavino, C., Murphy, E., Ogg, M., Smith, B.K., McAdams, S., 2011. Comparison of methods for collecting and modeling dissimilarity data: applications to complex sound stimuli. Multivariate Behavioral Research 46, 779–811.

Please cite this article as: White, A., et al., Problems of reliability and validity with similarity derived from category fluency. Psychiatry Research (2014), http://dx.doi.org/10.1016/j.psychres.2014.10.001

Henley, N.M., 1969. A psychological study of the semantics of animal terms. Journal of Verbal Learning and Verbal Behavior 8, 176–184.

Hutchison, K.A., Balota, D.A., 2003. Structure versus processing deficits in Alzheimer's disease, a matter of degree: a comment on Storms et al. (2003). Neuropsychology 17, 306–309.

Jarrold, C., 2003. What are the causes of reduced fit in scaling and clustering studies of semantic proximity data, and how else to measure them: a comment on Storms et al. (2003). Neuropsychology 17, 310–311.

Jarrold, C., Hartley, S.J., Phillips, C., Baddeley, A.D., 2000. Word fluency in Williams syndrome: evidence for unusual semantic organisation? Cognitive Neuropsychiatry 5, 293–319.

Lord, F.M., Novick, M.R., 1968. Statistical Theories of Mental Test Scores. Addison-Wesley, Reading, Massachusetts.

Milberg, W., McGlinchey, R., 2003. Taking the thumbs off the multidimensional scales in the debate on semantic memory and Alzheimer's disease: a comment on Storms et al. (2003). Neuropsychology 17, 312–314.

Moelter, S.T., Hill, S.K., Hughett, P., Gur, R.C., Gur, R.E., Ragland, J.D., 2005. Organization of semantic category exemplars in schizophrenia. Schizophrenia Research 78, 209–217.

Moelter, S.T., Hill, S.K., Ragland, J.D., Lunardelli, A., Gur, R.C., Gur, R.E., Moberg, P.J., 2001. Controlled and automatic processing during animal word list generation in schizophrenia. Neuropsychology 15, 502–509.

Ober, B.A., Shenaut, G.K., 1999. Well-organized conceptual domains in Alzheimer's disease. Journal of the International Neuropsychological Society 5, 676–684.

Ober, B.A., Shenaut, G.K., 2003. New directions in the study of semantic deficits: a comment on Storms et al. (2003). Neuropsychology 17, 315–317.

Paulsen, J.S., Romero, R., Chan, A., Davis, A.V., Heaton, R.K., Jeste, D.V., 1996. Impairment of the semantic network in schizophrenia. Psychiatry Research 63, 109–121.

Prescott, T.J., Newton, L.D., Mir, N.U., Woodruff, P.W., Parks, R.W., 2006. A new dissimilarity measure for finding semantic structure in category fluency data with implications for understanding memory organization in schizophrenia. Neuropsychology 20, 685–699.

Rogers, T.T., 2003. Is there madness in the method? A comment on Storms et al. (2003). Neuropsychology 17, 318–320.

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

Shallice, T., 1988. From Neuropsychology to Mental Structure. Cambridge University Press, Cambridge, England.

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

Storms, G., Dirikx, T., Saerens, J., Verstraeten, S., De Deyn, P.P., 2003b. On what we cannot learn from proximity data. Neuropsychology 17, 323–329.

Sumiyoshi, C., Ertugrul, A., Yagcioglu, A.E.A., Sumiyoshi, T., 2009. Semantic memory deficits based on category fluency performance in schizophrenia: similar impairment patterns of semantic organization across Turkish and Japanese patients. Psychiatry Research 167, 47–57.

Sumiyoshi, C., Matsui, M., Sumiyoshi, T., Yamashita, I., Sumiyoshi, S., Kurachi, M., 2001. Semantic structure in schizophrenia as assessed by the category fluency test: effect of verbal intelligence and age of onset. Psychiatry Research 105, 187–199.

Sumiyoshi, C., Sumiyoshi, T., Roy, A., Jayathilake, K., Meltzer, H.Y., 2006. Atypical antipsychotic drugs and organization of long-term semantic memory: multidimensional scaling and cluster analyses of category fluency performance in schizophrenia. International Journal of Neuropsychopharmacology 9, 677–683.

Sylvester, C.-Y.C., Shimamura, A.P., 2002. Evidence for intact semantic representations in patients with frontal lobe lesions. Neuropsychology 16, 197–207.

Takane, Y., 2003. Question hard, answer simply: a comment on Storms et al. (2003). Neuropsychology 17, 321–322.

Verheyen, S., Ameel, E., Storms, G., 2007. Determining the dimensionality in spatial representations of semantic concepts. Behavior Research Methods 39, 427–438.

Verheyen, S., Voorspoels, W., Vanpaemel, W., Storms, G., 2014. Caveats for the spatial arrangement method: comment on Hout, Goldinger, & Ferguson (2013) Journal of Experimental Psychology: General (under review).

Voorspoels, W., Storms, G., Longenecker, J., Verheyen, S., Weinberger, D.R., Elvevåg, B., 2014. Deriving semantic structure from category fluency: clustering techniques and their pitfalls. Cortex 55, 130–147.

Please cite this article as: White, A., et al., Problems of reliability and validity with similarity derived from category fluency. Psychiatry Research (2014), http://dx.doi.org/10.1016/j.psychres.2014.10.001