Metrics for Measuring Inter-item Dissimilarity in Fluency Data

Tony J. Prescott*, Lisa D. Newton, Nusrat U. Mir, Peter W. R. Woodruff, and Randolph W. Parks. University of Sheffield, UK.

This document is provided as an electronic appendix to the article "A new dissimilarity measure for finding semantic structure in category fluency data with implications for understanding memory organization in schizophrenia". This appendix contains two parts. In the first we derive a measure of the distance between two items in a fluency list that is appropriately normalized for list-length. This measure serves as the core element of the *mean cumulative frequency* (*mcf*) dissimilarity metric described in the article. In the second part we compare and contrast the *mcf* metric with the prevailing Chan et al. (1993) dissimilarity metric, referred to as the *dis* metric here and throughout the article.

1. Derivation of a Measure of Inter-item Distance based on Cumulative Frequencies

Our measure of normalized inter-item distance is based on an analysis of the frequency distribution with which any given inter-item distance *d* occurs within lists of different lengths. In the following, we first discuss this distribution, then derive the new distance measure by using calculations of cumulative frequency, and finally show how this metric can be weighted to take account of repeated items with a list.

The effect of list-length on the frequency of a given inter-item distance

Let *a* and *b* be the two items occurring at index positions (using the positive integers 1, 2, 3, …) i_{al} and i_{bl} in a given list *l* then we can calculate the 'raw' inter-item distance d_{abl} as the absolute difference of these values, i. e. $d_{abl} = |i_{al} - i_{bl}|$. If we calculate the inter-item distance for ever possible pair of items in a list of length *n,* then the frequency, $f(d,n)$, with which any specific distance *d* is observed, is given by the *triangular distribution* illustrated, for $2 \le n \le 10$, in Table 1. The table also shows the total number of item pairs for each list length, which we denote by $B(n)$.

*Author for correspondence. Address: Department of Psychology, University of Sheffield, Western Bank, Sheffield, S10 2TN, United Kingdom. Email: t.j.prescott@sheffield.ac.uk.

Table 1

Frequency $f(d,n)$ of 'Raw' Inter-item Distances *d* in Lists of Length *n*

Length	Inter-item distance (*d*)									Total
n	1	2	3	4	5	6	7	8	9	*B(n)*
2	1									1
3	2	1								3
4	3	2	1							6
5	4	3	2	1						10
6	5	4	3	2	1					15
7	6	5	4	3	2	1				21
8	7	6	5	4	3	2	1			28
9	8	7	6	5	4	3	2	1		36
10	9	8	7	6	5	4	3	2	1	45

From Table 1 we can see that the number of pairs obtained at a given distance, *d*, increases linearly with list length. At the same time, however, the proportion, or *relative frequency,* of pairs at a given distance, $f(d,n)/B(n)$, changes *non-linearly* with length. For example, in a list *n*= 5, 30% (3 out of 10) of inter-item distances are of length *d*= 2, but in a list twice this length (*n*= 10), item pairs with *d= 2* constitute just 18% of the combinations available (8 of 45). Differences in list-length therefore *do* matter, and in controlling for list length, we should take account of the dependence of relative frequency on both list length and inter-item distance.

Deriving a normalized distance measure based on cumulative frequency

To solve the problem of controlling for changes in relative frequency we now describe a measure of normalized inter-item distance, $D(d,n)$, that is based upon *cumulative frequency.*

For a list of length *n,* the cumulative frequency, $C(d,n)$, is the total number of item pairs with an inter-item distance less than or equal to *d,* i.e.

$$C(d,n) = \sum_{k=1}^{d} f(k,n).$$

So, for example, $C(3,6)$ can be obtained from Table 1 as the sum $f(1,6) + f(2,6) + f(3,6) = 5 + 4 + 3 = 12$. It is useful to be able to calculate this value directly without having to first generate the frequency distribution. We do this by noting that the total number of item pairs for a list of length n is $B(n) = n(n-1)/2$, and that the cumulative frequency, $C(d,n)$, is equal to the difference between the number of pairs in a list of length n and one of length $n-d$, i.e. $C(d,n) = B(n) - B(n-d)$. Finally, to obtain our normalized distance measure we take the average of $C(d,n)$ and $C(d-1,n)$ to find the centre of the frequency range for *d*, and divide by the total number of item pairs $B(n)$, hence

$$D(d,n) = \tfrac{1}{2}\left(C(d,n) + C(d-1,n)\right)/B(n)$$

which simplifies to

$$D(d,n) = (2dn - d^2 - n)/n(n-1). \qquad (1)$$

Figure 1 provides a graphical illustration of the proposed measure for a list containing seven items showing that, for any given *d*, the normalized measure exactly bisects the range of the relevant frequency distribution. So, for example, $D(2,7) = 0.40$ is the mid-point of the cumulative frequency range 1–2, which is the average position, in this distribution, of all item pairs with *d*= 2.

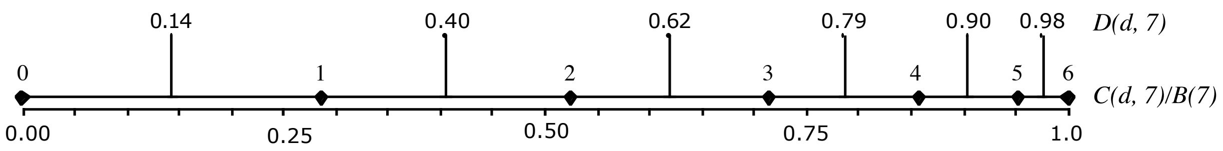

Figure 1. An illustration of the normalized inter-item distance measure, $D(d,n)$, for lists of length 7. The figure shows the cumulative frequency for 'raw' inter-items distances of 0–6, and, above this, the value of $D(d,7)$ for each possible observed inter-item distance. For any given value of *d,* $D(d,n)$ exactly bisect the range of the relevant frequency distribution.

In order to apply the normalized distance measure derived above to specific items in fluency lists we introduce indexing with respect to items *a* and *b* in list *l,* of length n_l, giving

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

Dealing with repeated items within lists

The above analysis treats every item pair in a fluency list as a separate entity to be assigned its own distance measure, we now describe how this method can be extended to deal appropriately with lists in which one or more items are repeated. If either of the items *a* or *b* occurs more than once in *l* we select our 'raw' measure of inter-item distance $\hat{d}_{abl}$ to be the smallest of all such distances, i.e. $\hat{d}_{abl} = \min_{\forall a,b \in l} d_{abl}$. Selecting the smallest inter-item distance for a repeated item pair does not, however, entirely resolve the problem as repetitions also significantly increase the frequency of short inter-item distances relative to long ones. Rather than determine an exact solution for such situations, we have developed an approximation, based on equation 1, through an analysis of the relevant frequency distributions. Specifically, let o_{al} and o_{bl} be the number of occurrences of items *a* and *b* respectively in list *l*, then a measure of normalized inter-item distance that adjusts for repeated items can be calculated using

$$\hat{D}_{abl} = \left[D(\hat{d}_{abl}, m_{abl})\right]^{\lambda_{abl}}, \tag{3}$$

where $m_{abl} = n_l - (o_{al} + o_{bl} - 2)$ is the length of the list less the number of repetitions of *a* and *b*, and $\lambda_{abl} = (o_{al}o_{bl})^{-2/3}$. Here λ_{abl} provides an exponential scaling that increases the value of the measure to compensate for the greater frequency of short inter-item distances in a list containing repeated items. The procedure used to derive this approximation was (i) to generate all possible lists of lengths between 2 and 20 containing combinations of $o_a \in \{1,2\}$ and $o_b \in \{1,2,3\}$ occurrences of items *a* and *b*, (ii) to use this data to generate target cumulative frequency distributions, and (iii) to select the exponent of the scaling factor λ_{abl} so as to minimise the sum squared error between the approximation (equation 3) and the set of target inter-item distance values generated directly from this frequency data.

2. Comparison of the *mcf* and *dis* measures of inter-item dissimilarity

We next compare the *mcf* metric with a dissimilarity measure proposed by Chan et al. (1993) and referred to here as the *dis* metric. The two metrics can be written as

$$\mathrm{mcf}(G,a,a)=0,\quad \mathrm{mcf}(G,a,b)=\frac{1}{T_{Gab}}\left(\sum_{l\in G;a,b\in l}\hat{D}_{abl}\right), \tag{4}$$

$$\mathrm{dis}(G,a,a)=0,\quad \mathrm{dis}(G,a,b)=\frac{N_g}{T_{Gab}^2}\left(\sum_{l\in G;a,b\in l}\frac{d_{abl}}{n_l}\right). \tag{5}$$

where G is the participant group, N_G is the total number of lists generated by G, and T_{Gab} the number of such lists containing both *a* and *b*. It is evident from the above, that the *dis* and *mcf* metrics differ both in the way that distances are calculated *within* the list of a single participant, and in how they are combined *across* participants. We briefly consider both sources of difference below.

First, the *dis* measure uses division by list length (*n)* to generate normalized distance measure for any pair of items in a given list. In some respects this is a fair approximation to the algorithm we have derived from our analysis of the underlying frequency distribution. For instance, equation 1 reduces exactly to *1/n* for the particular case of *d=1.* For values of *d* greater than 1, however, division by length significantly under-estimates the frequency-based measure. This is illustrated for two example list lengths in Figure 2.

Second, on the issue of combining estimates across participants, we can see from the above comparison of the two metrics that the *dis* metric does not take a simple average of normalized inter-item distance across all participants (this would be division by T_{Gab} as in the *mcf* metric) but instead weights this average by N_G/T_{Gab} (giving N_G/T_{Gab}^2 overall). In their description of the algorithm, Chan et al. (1993) justify this weighting by alluding to a goal of compensating for differences in the relative frequency of item pairs across the participant group, however, they do not explain how their algorithm serves this purpose. We will therefore briefly explore this idea and consider how such compensation might be best achieved.

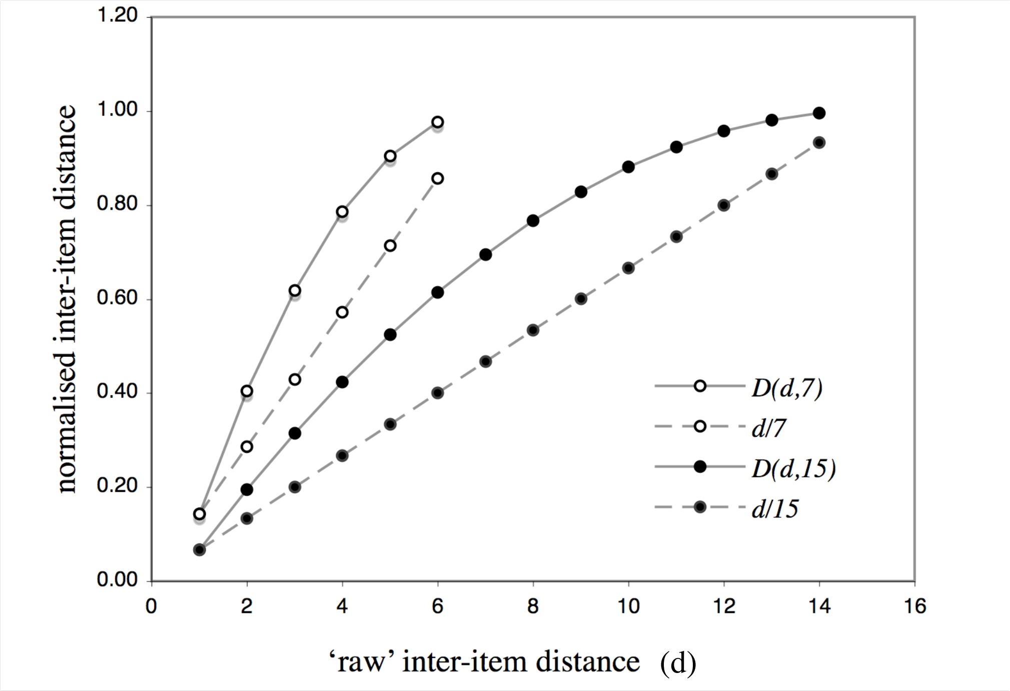

Figure 2. The mapping from 'raw' inter-item distances, *d*, into the normalized measure, $D(d,n)$ (Equation 2), for two example list lengths (*n*= 7, 15), compared with one calculated using d/n (as in Chan et al., 1993). The graph indicates that division by list length provides a first-order approximation to the frequency-based mapping, that is accurate at *d=1* but otherwise under-estimates the latter across most of its range.

In a previous article (Crowe & Prescott, 2003), we have argued that, across a set of fluency lists, differences in the frequency with which pairs of items occur together can provide a useful indicator of item similarity. This idea derives from the reasonable assumption that items that are closely related in semantic memory will be named together in fluency lists more often than items that are distantly related (see also Schwartz and Baldo, 2001, for a method of calculating item similarity based on this principle). It seems likely, then, that the additional weighting in equation 5 is intended to serve this purpose. Indeed, N_G/T_{Gab}, as a multiplier of the within-list dissimilarity measure, certainly acts in the right direction—its value is smallest (reflecting item similarity) for items that co-occur with high frequency and largest (reflecting dissimilarity) for items that co-occur with low frequency. However, there are two significant disadvantages to using the frequency of item co-occurrence in this way. The first problem is that N_G/T_{Gab} is strongly negatively correlated with item *production frequency*. That is, words that are named often in lists will also co-occur frequently and will therefore be treated as more semantically related. The second problem is that by using the inverse of T_{Gab} the weighting applied is non-linear; this serves to further exaggerate the increased proximity of high frequency items and decreased proximity of low ones. For instance an item pair that occurs once in 10 pairings will have a weighting of 10, one that occurs twice, a weighting of 5, and one that occurs five times a weighting of just 2. Hence this algorithm makes it very difficult to obtain a low dissimilarity rating for a low frequency pair. Note that it is possible to devise a measure that combines information about item co-occurrence with average inter-item distance whilst avoiding these pitfalls. We have described such a combined measure—based on comparing the *observed* frequency of an item pair with its *expected* value—in Crowe and Prescott (2003), where we suggest that this may provide a useful indicator of dissimilarity when studying populations, such as young children, that generate very short fluency lists. In the current article, however, we have investigated lists from adult patients and controls that are of sufficient length to justify using a measure based on inter-item distance alone.

References

Chan, A. S., Butters, N., Paulsen, J. S., Salmon, D. P., Swenson, M. R., & Maloney, L. T. (1993). An assessment of the semantic network in patients with Alzheimers- disease. *Journal of Cognitive Neuroscience, 5*(2), 254-261.

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

Schwartz, S., & Baldo, J. (2001). Distinct patterns of word retrieval in right and left frontal lobe patients: a multidimensional perspective. *Neuropsychologia*, *39*, 1209-1217.