*Journal of the International Neuropsychological Society* (2002), **8**, 425–435.

DOI: 10.1017.S1355617701020227

# New semantic and serial clustering indices for the California Verbal Learning Test–Second Edition: Background, rationale, and formulae

JOHN L. STRICKER,[1,2] GREGORY G. BROWN,[1,3] JOHN WIXTED,[4]
JULIANA V. BALDO,[5] AND DEAN C. DELIS[1,3]

[1]Psychology Service, Veterans Administration San Diego Healthcare System
[2]Joint Doctoral Program in Clinical Psychology, San Diego State University/University of California San Diego
[3]Psychiatry Department, University of California San Diego
[4]Psychology Department University of California San Diego
[5]Center for Aphasia and Related Disorders, VA Northern California Health Care System

(Received June 27, 2000; Revised April 4, 2001; Accepted April 10, 2001)

**Abstract**

The original California Verbal Learning Test (CVLT) employed a semantic clustering index that used the words recalled during a given trial as the baseline for calculating expected values of chance clustering (recall-based expectancy). Although commonly used in cognitive psychology, clustering indices that use *recall-based* calculations of expectancy are implied by the assumption that organizational processes do not occur until *after* words are retrieved from memory. This assumption contradicts the generally held assumptions among neuropsychologists that (1) organization is an antecedent to recall, and (2) increases in the use of organizational strategies will result in better recall performance. After reviewing a brief history of clustering metrics, we used Monte Carlo simulations, informative examples, and patient data to examine clustering indices that use the *word list* as a baseline for calculating expectancy and propose these list-based expectancy measures as a refinement of the clustering indices used on the original CVLT. These indices are used on the recently published CVLT–II. (*JINS*, 2002, *8*, 425–435.)

**Keywords:** Semantic clustering, Memory organization, California Verbal Learning Test

## INTRODUCTION

The California Verbal Learning Test (CVLT; Delis et al., 1987) incorporated principles from cognitive psychology to provide multiple indices of an individual's ability to learn and remember verbal material. The CVLT is a process-oriented memory task that, in addition to assessing many aspects of learning and memory, also evaluates encoding strategies that may have been used by subjects. The original CVLT identified two measures of learning strategies: semantic clustering and serial clustering.

Recently, we have identified a controversy involving the semantic clustering formula that traditionally has been used in cognitive research and in the original CVLT. Before going into the details of the issue, we will start with a simple demonstration of how the original CVLT's semantic clustering index would score two different recall patterns.

The CVLT's target list consists of 16 words, with four words from each of four categories. The most obvious behavioral evidence of the use of an organizational strategy is the extent to which organization is demonstrated in the recall pattern. In the case of semantic clustering, the use of categories to assist in recall could be measured by the extent to which an individual consecutively recalls words from the same category. If a category cluster is defined as a pair of consecutively recalled words from the same category, the more clusters of words from the same category observed in the subject's recall order, the more likely it is that a subject is using semantic information to assist in recall. Consider the following two recall patterns: (1) "grapes, apricots, pliers, wrench," (2) "tangerines, plums, grapes, apricots." Intuitively, it might seem that Example 2 would receive a higher clustering score than Example 1 because three semantic clusters are present in Example 2, whereas only two

Reprint requests to: Gregory G. Brown, Psychology Service (116B), Veterans Administration San Diego Healthcare System, 3350 La Jolla Village Drive, San Diego, CA 92161. E-mail: gbrown@ucsd.edu

**Table 1.** Comparing the original CVLT Semantic Clustering Index with $LBC_{sem}$ and ARC

| Example | Category recall order | Words recalled | Observed clusters | Original CVLT semantic ratio | $LBC_{sem}$ | ARC |
|---|---|---|---|---|---|---|
| 1 | a, a, b, b | 4 | 2 | 2.00 | 1.40 | 1.00 |
| 2 | a, a, a, a | 4 | 3 | 1.00 | 2.40 | Undefined |
| 3 | a, a, a, a, b, b, b, b | 8 | 6 | 2.00 | 4.62 | 1.00 |
| 4 | a, a, b, b, c, c, d, d | 8 | 4 | 4.00 | 2.62 | 1.00 |
| 5 | a, a, a, a, b, b, b, b, c, c, c, c | 12 | 9 | 3.00 | 6.80 | 1.00 |
| 6 | a, a, a, b, b, b, c, c, c, d, d, d | 12 | 8 | 4.00 | 5.80 | 1.00 |
| 7 | a, a, b, b, c, d, c, d | 8 | 2 | 2.00 | 0.62 | 0.33 |
| 8 | a, a, a, a, c, d, c, d | 8 | 3 | 1.50 | 1.62 | 0.33 |
| 9 | a, a, b, b, a, b, a, b | 8 | 2 | 0.67 | 0.62 | 0.11 |
| 10 | a, a, a, b, b, b, c, c, c, d, d, d, a, b, c, d | 16 | 8 | 2.67 | 4.99 | 0.55 |

*Note*. Each letter (a–d) represents a word belonging to one of the four categories on List A of the CVLT.

clusters are present in Example 1. *However, the original CVLT's semantic clustering score for Example 1 would be twice as large as Example 2*.

In order to explain why the original CVLT's semantic clustering index behaves the way it does, and to provide a rationale for the use of the semantic and serial clustering indices that are incorporated into the CVLT–II, we first briefly review the historical and conceptual background of the semantic clustering index used by the CVLT. We then discuss the theoretical assumptions underlying the method of calculating semantic clustering that was adopted by the original CVLT and other neuropsychological studies of category clustering. We next describe the methods used to calculate the original CVLT clustering indices and alternative indices, including those used by the CVLT–II. Finally, using Monte Carlo simulations, informative examples and patient data, we compare the original CVLT's semantic and serial clustering with the new clustering indices in the CVLT–II.

## A Brief History of the Role of Organization in Verbal Learning

Since the earliest writings on verbal learning theory and research, theorists have argued that greater organization of the material to be learned should be associated with increased recall performance (James, 1890; Hilgard, 1948; Mandler, 1967). Miller's (1956) classic theory of limited capacity in short-term information storage (the famous 7 ± 2 items) prompted psychologists to theorize that the organization of material during the learning process might compensate for such memory limitations (see Mandler, 1967, for a review).

Verbal learning theorists have proposed several types of organizational strategies, and numerous ways to measure each type. Tulving (1968) identified two broad categories of organization: primary and secondary organization. Primary organization is based solely on an individual's experience with the word list as it is presented. Serial-position effects, such as primacy and recency, are examples of primary organization. In contrast, secondary organization refers to organization based upon an individual's previous experience with the presented words. Secondary organization can occur by utilizing common associations between the words used on a list (e.g., associating words from a categorized list by category) or by grouping words together in a seemingly arbitrary manner that is presumably meaningful to the individual (e.g., subjective organization).

The superiority of secondary organization over primary organization as a learning strategy, in particular the use of categories to remember a word list, has often been demonstrated (Craik, 1981; Mandler, 1967). Given the evidence, Tulving (1968) suggested that primary organization is the result of short-term processes, and is not reflective of learning. Secondary organization, on the other hand, may be more related to learning. The experimental findings regarding the role of organizational strategies are of interest to clinical neuropsychologists who study memory dysfunction. Memory deficits observed in clinical populations may be due to deficits in secondary organization and/or to the overreliance upon primary organization.

## Overview of Clustering Indices: Observed and Expected Values

Several researchers developed formulas to compute the amount of organization that has occurred based upon the content of a subject's recall. The method by which observed clustering has been scored is relatively uncontroversial. For most investigators, an observed category cluster occurs anytime that two words in adjacent recall positions are members of the same category.[1]

As a subject recalls more words, it becomes increasingly likely that clusters of the same category will appear together

[1] Although one critic has argued that the effective use of an organizational strategy does not necessarily imply adjacency in recall (Buschke, 1975).

during recall simply by chance. As a consequence, even if the subject were not using an organizational strategy, the amount of observed clustering would increase as recall increases. To control for observed clusters that may occur by chance, clustering indices adjust the amount of observed clustering by a value that would be expected if the subject recalled words without organizing them. In contrast to the calculation of observed clustering, there has been controversy regarding how chance-expected values should be calculated (see, e.g., Colle, 1972; Shuell, 1969; Pellegrino & Hubert, 1982).

One important issue in determining how chance-expected clustering values should be calculated involves the choice of which baseline set of words should be used to calculate expected values. Generally, one of two baselines have been chosen: (1) the words that have been recalled by the subject (recall-based expectancy); or (2) the entire study list (list-based expectancy). The authors of the CVLT adopted a semantic clustering index described by Shuell (1969), who in turn used Bousfield's (1953) recall-based expectancy formula. Although Bousfield and his colleagues initially used list-based expectancy for clustering measures that required a calculation of expectancy (the IR measures; see Bousfield & Cohen, 1952; Bousfield, 1953; Bousfield & Cohen, 1955; Bousfield et al., 1964), Bousfield and Puff (1964) later dismissed these methods and introduced a measure (ITR) based on the following two related assumptions. First, when an individual undertakes recall of a list, a certain number of items from the list are not available to that individual and therefore have no effect on the degree of clustering. Second, at any stage of recall, all the words available to be recalled are equally available and are chosen without replacement. Note that some words might have been encoded and retained, yet not retrieved. These two assumptions are also embedded in the original CVLT's semantic clustering index.

When the ITR measure was first introduced, Bousfield and his colleagues (Bousfield et al., 1964; Bousfield & Bousfield, 1966) justified the first assumption by referring to Underwood and Schulz's (1960) theory of meaningfulness in verbal learning. In this theory, the frequency of exposure to a word is considered to be the "fundamental antecedent" to subsidiary processes—such as semantic and serial clustering. Consequently, the extent to which a person is able to recall a group of words is dependent upon frequency of exposure.

After 1966, many researchers adopted the assumption that, for semantic clustering in particular, organization of recall occurred *after* words were retrieved, and that any calculations of chance-expectancy scores should use as a baseline the words recalled by an individual rather than the list to be learned. As other determinants of recall besides meaningfulness became identified, Underwood and Schulz's (1960) theory became less tenable and was no longer used to support the recall-based expectancy assumptions. As Underwood and Schulz's (1960) theory became less often cited, the justification of recall-based expectancy corrections splintered into a variety of rationales. For example, Shuell (1969) argued that list-based expectancy methods of calculating chance clustering unreasonably assumed that all words presented to the subject are equally available for recall. Since all of the words presented to the subject were not available to the subject, as demonstrated by the widely-observed serial position effect, Shuell (1969) opted for the more limited assumption that category clustering occurred only for those items that were retrieved during recall. Pellegrino and Hubert (1982) also defended recall-based expectancy by stating that if we were to use the word list as the baseline, it would reflect two sources of information: that of content (which words were recalled) and that of structure (how the words were organized). According to Pellegrino and Hubert (1982), using the words that were recalled by the subject as the baseline would allow us to measure structure separately from content.

Although different explanations are provided, both Shuell's (1969) and Pellegrino and Hubert's (1982) rationales, like the theory stated by Underwood and Shultz (1960), is most clearly reconciled with a theory of learning where organizational processes do not affect which items are recalled. However, in many cognitive and neuropsychological studies, investigators have drawn conclusions about the organizational strategies used during encoding or retrieval from clustering indices using recall-based expectancy to correct for chance clustering (Crosson et al., 1989; Massman et al., 1990; Van Spaendonck et al., 1996). Like many psychologists influenced by cognitive researchers (e.g. Mandler, 1967), the authors of the CVLT proposed that the "semantic clustering strategy typically results in the most effective *encoding* into long-term memory" (Delis et al., 1987, p. 17; italics added). As we will show below using individual cases, the recall based expectancy measure that is used in the CVLT's semantic clustering index produces scores that are inconsistent with the idea that the joint relationship between organization and recall is being measured.

## CLUSTERING INDICES OF THE ORIGINAL CVLT

The CVLT begins with the presentation of a 16 word "shopping list." This list consists of words containing four words each from four semantic categories. The words are organized in the list such that no two words from the same category are in adjacent positions. Subjects are asked to recall as many words from the shopping list as they can after the list is presented. The first phase of the CVLT consists of five presentations of the word list; each followed by immediate recall. The word list is presented in the same order for all five learning trials. The type of encoding strategy used by the subjects to learn the word list is inferred by examining the order in which subjects recall words on each of the five trials. Semantic and serial clustering scores are calculated for each recall and then averaged across all five trials.

## The CVLT's Semantic Clustering Index

Like most clustering measures, the CVLT's semantic clustering index can be broken into two components: the observed and expected values. Because a category cluster occurs anytime that two words in adjacent recall positions are members of the same category, the maximum number of observed clusters on the CVLT is 12. The second component of the CVLT's semantic clustering index is the number of category clusters expected by random recall. The expected semantic clustering score for each trial described in original CVLT manual (Delis et al. 1987) is equivalent to Bousfield and Bousfield's (1966) formula:

$$\text{Expected semantic clustering for each trial} = \sum_{i=1}^{4} \frac{[n_i(n_i - 1)]}{r} \tag{1}$$

where $i$ = category type, $n_i$ = the number of correct words recalled from category i on a given trial, and $r$ = the total number of words recalled on a given trial, including intrusions and repetitions. The maximum value of the expected semantic clustering index is 3, which occurs when all 16 words are recalled (such that $n_i = 4$ for $i = 1$ to 4, and $N = 16$).

Although the Bousfields subtracted the expected score from the observed clustering score, Shuell (1969) observed that Bousfields' difference score did not correct for differences in the maximum amount of clustering when disparities in the number of words recalled occurred across conditions or groups to be compared. To adjust for differences in maximum recall, Shuell (1969) recommended that observed semantic clustering be divided by the expected score to yield a ratio. The authors of the original CVLT have adopted the Bousfields' formula, but they followed Shuell's (1969) advice and divided the observed clusters by the expected, rather than subtracting. This ratio is then interpreted as a multiple of the degree to which observed clustering was expected to occur by chance. For example, if an examinee's semantic clustering ratio was 2.5, then the examinee would have used clustering 2.5 times that expected by chance. The maximum value of the semantic clustering index is 4.

## The CVLT's Serial Clustering Index

The serial clustering index of the original CVLT resembles the semantic clustering index in that both are ratios of observed over expected values. The observed serial clustering score is calculated in much the same way as the observed semantic clustering score: each time two words are recalled in the same sequence in which they are presented on the list, the observed score increases by 1.

However, the expected serial clustering score is not analogous to the expected semantic clustering score. The expected serial clustering score is calculated by using a power function generated from a Monte Carlo study which examined recall lists of varying length (unpublished data by Fridlund & Delis, 1983; cited in Delis et al., 1987). The expected clustering score is calculated using the following formula (from Delis et al., 1987):

$$\text{Expected serial clustering for trial } i = (0.135 \times \#C_i^{0.62}) - 0.135 \tag{2}$$

where $\#C_i$ = number of correct words recalled on trial $i$. Expected serial clustering values range from zero (when $\#C_i$ is 1) to 0.618 (when $\#C_i$ is 16). The serial clustering ratio ranges from a minimum of zero (16 words recalled with zero observed clusters) to a maximum of 24.266 (16 words recalled with 15 observed clusters).

# ALTERNATIVE CLUSTERING INDICES

## Adjusted Ratio of Clustering

Shuell (1969) proposed forming a ratio of observed to expected clustering to correct for differences in maximum clustering associated with different recall totals. However, as will be shown below, this ratio does not adequately correct for different recall maxima. Another clustering measure based upon recall-based expectancy, the *adjusted ratio of clustering* (ARC: Roenker et al., 1971), does adjust for varying maximum clustering values. The formula for ARC is as follows:

$$\text{ARC} = \frac{\text{OBS} - \text{EXP}}{\text{MAX} - \text{EXP}} \tag{3}$$

where EXP is calculated in the same manner as it is for the CVLT's semantic clustering ratio (see Equation 1), and MAX is calculated by subtracting the number of different categories that were recalled from the total number of words recalled.[2] Note that, like its expected value, ARC's MAX is calculated using the subject's recall, not the word list. Thus, like Bousfield's (1953) formula, ARC's MAX and its calculation of chance-expectancy are recall-based.

[2]The calculation of expected values for ARC is often represented as a more compact formulation of Equation 1:

$$\text{ARC EXP} = \frac{\sum n_i^2}{N} - 1 \tag{7}$$

To show how these two equations are equivalent, Equation 1 can be expanded to:

$$\frac{\sum n_i^2}{N} - \frac{\sum n_i}{N} \tag{8}$$

In the later half of Equation 8, the sum of the number of all instances of each category ($\sum n_i$) will always equal the total number of words recalled ($N$), so the division of the two will always equal 1. Thus, Equation 7 is equal to Equation 8, and Equation 7 is equal to Equation 1. However, if intrusions and repetitions are included in the calculation of $N$, then Equation 1 and Equation 7 will not be equivalent when any intrusions or repetitions are present.

ARC has a maximum of 1.0. Although negative values may be obtained with ARC, it is difficult to interpret the meaning of these values given that ARC attempts to provide a percentage of organization score (e.g., given a recall of 5, what would it mean if $-10\%$ of that recall was organized?).

## List-Based Semantic Clustering Index

Instead of using recall-based expectancy to control for variations in chance recall due to different levels of recall, an alternative approach would be to use list-based expectancy. Frender and Doubilet (1974) developed the following list-based expectancy formula to correct for chance recall:

$$\text{EXP}_{\text{sem}} \text{ for a given trial} = \frac{[(r-1)(m-1)]}{N_{\text{L}}-1} \tag{4}$$

where $r$ = the number of correct words recalled on that trial, $m$ is the number of members of each semantic category on the original list (category size is assumed to be equal across categories), and $N_{\text{L}}$ = the total number of words on the original list. When applied to the original CVLT, or the CVLT–II, both of which have 16 words per list with four categories represented ($N_{\text{L}} = 16$, $m = 4$), the formula simplifies to

$$\text{EXP}_{\text{sem}} = [(r-1)/5] \tag{5}$$

Frender and Doubilet (1974) did not specify their rationale for this equation. However, we independently developed this formula as well as its mathematical proof (see Appendix A). Like the recall-based expectancy formula, the maximum list-based expectancy value is 3 for the CVLT. However, unlike the recall-based expectancy formula, *the expectancy value is based on the categories represented on the original list, not the words recalled*. Also, unlike the CVLT's semantic clustering index, intrusions and repetitions are not included in the calculation of expectancy.

The clustering index, which we will refer to as the List-Based Clustering Index (LBC), is calculated by subtracting $\text{EXP}_{\text{sem}}$ (equation 5) from the number of observed clusters for each trial:

$$\text{LBC}_{\text{sem}} = \text{OBS}_{\text{sem}} - \text{EXP}_{\text{sem}} \tag{6}$$

We have chosen to use an equation consisting of the difference between the observed and expected values as opposed to a ratio so that negative values of the index would continue to be meaningful. LBC can be interpreted as how many observed clusters have occurred in an individual's recall beyond what would be expected if the individual randomly recalled words from the list to be learned. For the CVLT, LBC has a maximum of 9 (when all 16 words are recalled and organized by category) and a minimum of $-3$. Large negative LBC values indicate that other forms of organization besides semantic clustering are likely to be occurring. For example, an individual who recalled all 16 CVLT words in their original presentation order would obtain the most negative LBC score possible.

## List Based Serial Clustering Index

Using list-based expectancy, an expected score for serial clustering can also be calculated. For the CVLT or CVLT–II, an expected score for a single trial can be calculated by using the following formula (see Appendix B for the general formula and mathematical proof):

$$\text{EXP}_{\text{ser}} = (r-1)/N_{\text{L}} \tag{7}$$

where, as before, $r$ is the number or words recalled on a given trial and $N_{\text{L}}$ is the number of words on the list. For the CVLT and CVLT–II, $N_{\text{L}}$ is 16. Note that the maximum expected value (when 16 words are recalled so that $r = 16$) is less than 1 ($\text{EXP}_{\text{ser}} = 15/16$). This is because serial clustering would not be very likely to occur by chance, even if an individual were to randomly recall the entire word list.

A serial LBC index can be calculated by subtracting the expected value generated by Equation 7 from the number of observed serial clusters (as in Equation 6).

$$\text{LBC}_{\text{ser}} = \text{OBS}_{\text{ser}} - \text{EXP}_{\text{ser}} \tag{8}$$

# EVALUATING THE CLUSTERING INDICES

At a minimum, a chance-corrected clustering index should meet two criteria: (1) When clustering is random, the index should yield a constant value indicating that no clustering has occurred above that expected by chance, regardless of the total number of words recalled; and (2) the index should provide meaningful scores that reflect degree of organization when clustering is not random. We evaluate the original CVLT clustering indices in terms of these criteria using Monte Carlo simulations to study random clustering and hypothetical examples to study each index's measurement characteristics when gauging non-random organization.

## The Original CVLT Semantic Clustering Index

### *First criterion*

In order to test the CVLT semantic clustering index's sensitivity to variations in total recall when clustering was random, we programmed a recall simulation that randomly sampled without replacement from the CVLT's word list. Scoring of semantic and serial clustering in the simulation was crosschecked for accuracy with the CVLT's computerized scoring program[3] (Fridlund & Delis, 1987). As Figure 1 indicates, the resulting simulation average was a CVLT semantic clustering index of 1.0. Thus, mean semantic clustering was independent of recall total, indicating that the original CVLT's semantic clustering index does adjust for variation in total recall when clustering is random.

### *Second criterion*

Although the original CVLT's semantic clustering index adequately adjusts expected clustering for variation in total recall when clustering is random, the index is difficult to interpret when nonrandom clustering occurs. We will illustrate this point by using several possible examples of free recall from the CVLT's word list. Referring back to Table 1, Examples 1 through 8 illustrate that the original CVLT semantic clustering index had a superordinate preference; it favored the number of categories recalled over the number of words recalled in each category. Comparing Examples 1 and 2 in Table 1, note that the same number of words was recalled. Because all four words are from the same category in Example 2, there are no other possible category combinations in which those four words can be recalled (remember, the words recalled represent the baseline), hence the semantic ratio is 1. Recall that a semantic clustering value of one is also expected from random clustering. Yet from the perspective from the list to be learned, it is unlikely that four consecutively recalled words from the same semantic category would have been recalled by chance ($p \sim .002$ for the CVLT list).

Examples 3 and 4 and Examples 5 and 6 in Table 1 show that the superordinate preference is not specific to recalling four words from one category. Even when more words are recalled, the semantic clustering ratio still awards higher scores to recall patterns containing smaller, yet more numerous category clusters. Examples 7 and 8 illustrate that the original CVLT semantic cluster ratio demonstrates a bias towards smaller and more numerous clusters when additional words are recalled that do not belong to the categories that were clustered.

Examples 1 (a, a, b, b) and 3 (a, a, a, a, b, b, b, b) in Table 1 highlight a core feature of recall-based expectancy measures. Because only two categories are recalled in both of these examples, the original CVLT's semantic clustering index awards the same score. However, if a clustering index is intended to measure the relationship between recall and organization, then the index should increase as clustered recall increases.

Another problem with the original CVLT semantic clustering measure is the use of a ratio index to correct for

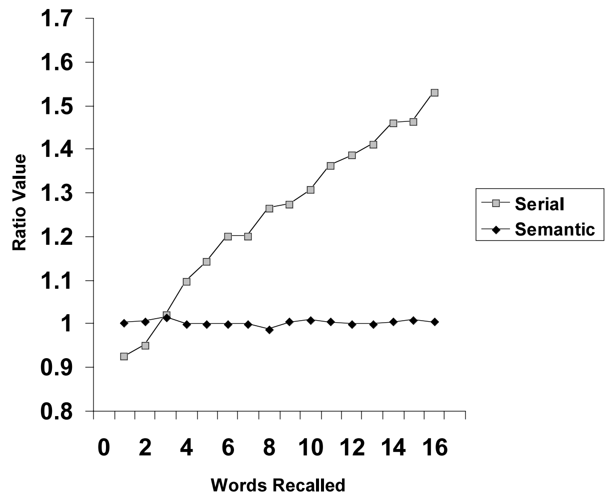


**Fig. 1.** Serial and semantic clustering indices calculated from simulations of random recall.

chance performance. When the observed amount of clustering does not equal zero, the clustering ratio is interpreted as the extent to which the amount of observed clustering differed from the amount of expected clustering (assuming that the measures of clustering are valid). However, when the amount of observed clustering does equal zero, the ratio can no longer be interpreted as a measure of the extent of clustering because the expected value no longer affects the ratio. If a subject's observed clustering score was zero, then the semantic clustering ratio would equal zero regardless of how much semantic clustering would be expected by chance. Hence, when the observed value equals zero, the clustering index can be interpreted only as indicating that no observed clustering was found. Yet, the extent to which the subject recalled fewer clusters than expected by chance might be clinically informative (for example, it may indicate that some other organizational strategy is being utilized).

Comparing the CVLT index with ARC across the first six examples in Table 1 highlights a final issue with the original CVLT clustering index. As can be seen in the recall in Table 1 (with the exception of Example 2, where ARC is undefined), ARC awards a maximum clustering score when a maximum amount of organization within the set of recalled words has occurred. In contrast to the ARC scores, the original CVLT's semantic clustering ratio less consistently determined that these examples were maximially organized. Even within the framework of recall-based indices, the original CVLT semantic clustering index does not consistently identify maximum clustering.

## The List-Based Semantic Clustering Index

As described above, the semantic *list-based clustering index* ($LBC_{sem}$) uses the target word list as a baseline for calculating chance expected values. As Table 3 shows, the number of chance clusters predicted from the list-based ex-

[3]This process revealed an error in the CVLT's computerized scoring program (Fridlund & Delis, 1987). During the first five trials of List A, whenever the word "slacks" is followed by at least one error (either an intrusion, repetition, or some combination) and then the word "drill," the program will incorrectly score this as serial clustering. If an error does not occur between "slacks" and "drill," then the program correctly does not score the pair as a serial cluster.

pectancy formula (Equation 5) is nearly identical to the number of clusters observed in our Monte Carlo simulation of random recall, regardless of the number of words recalled. Thus when recall is random, the expected $LBC_{sem}$ is zero and is independent of the number of words recalled. Table 1 illustrates the properties of $LBC_{sem}$ and compares this index with the original CVLT semantic clustering ratio. As demonstrated in Table 1, $LBC_{sem}$ is designed to increase as both recall *and* clustering increase. While the original CVLT's Semantic Ratio awards the maximum clustering score for Examples 4 and 6 in Table 1, $LBC_{sem}$ awards higher scores as the number of observed clusters increase beyond what would be expected by chance. Furthermore, $LBC_{sem}$ awards more points for longer clusters of fewer categories, while the CVLT's semantic clustering index rewards more points for smaller clusters of more numerous categories (compare the CVLT Ratio with $LBC_{sem}$ scores in Examples 1 and 2 and in Examples 3 and 4 in Table 1). $LBC_{sem}$ awards more points for longer clusters of the same category because when randomly sampling from the word list, longer clusters of the same category would be less likely to occur by chance than smaller strings of different categories.

## Rational Comparison of $LBC_{sem}$ to ARC and to the Original CVLT Semantic Clustering Index

ARC does not show the same superordinate preference that the original CVLT does, and may be considered an alternative to the original CVLT index. Schmidt (1997) has demonstrated that ARC yields a consistent value of zero with random clustering across different recall totals. As Table 1 indicates, ARC assigns similar values whenever increasing organization is balanced by increases in total recall. In contrast, the co-occurrence of longer clusters and greater recall invariably yields larger $LBC_{sem}$ values. Compare ARC and $LBC_{sem}$ values for Examples 1 and 3 in Table 1.

Because list-based expectancy relies on a fixed baseline represented by the category structure of the target list, the maximum cluster value is fixed, and the number of words recalled determines the number of clusters expected by chance. As the same number of words was recalled in Examples 3 and 4, the expected chance cluster value was the same. However, in Example 3 more clusters were recalled, producing a higher $LBC_{sem}$ value. In contrast, because ARC relies upon recall-based expectancy, expected values depend upon the amount of organization that is possible given the categories and number of words that were recalled. From the ARC perspective, the expected number of clusters based on chance is greater in Example 3 than in Example 4; the differences in expected chance clusters balances the differences in the number of observed clusters.

Several researchers have justified using ARC as opposed to other possible indices based on the results of Monte Carlo simulations where non-zero levels of clustering are crossed with different levels of recall (Murphy 1979; Schmidt, 1997). In these simulations, where clustering and recall levels are independently manipulated, ARC typically demonstrates the least amount of variance due to recall, while accurately mirroring differences in organizational level (see Murphy, 1979 for more details). Note, however, that these simulations make the same assumption that recall-based indices make: that increases in recall are not related to increases in organization. Given that this assumption is built into these simulations, and also into ARC, it is not surprising that ARC performs so well in these simulations. However, because the simulations independently manipulate recall and clustering, they fail to reflect memory-processing theories that assume that recall *is* related to clustering, and thus they are not adequate tests of list-based clustering indices.

In sum, ARC is a desirable measure of the amount of clustering given the number of words and categories recalled. However, if the aim of a clustering index is to measure the extent to which individuals use category organization to learn and retrieve a structured list, the co-occurrence of increasing semantic clustering and greater total recall should progressively move the index closer to maximum clustering. Based on this criterion, $LBC_{sem}$ is the preferred measure.

## Empirical Comparison of $LBC_{sem}$ to ARC and to the CVLT's Original Semantic Clustering Index

To obtain preliminary information about the performance of the original semantic clustering index, $LBC_{sem}$, and ARC, we compared a group of ten patients with frontal lobe lesions with 10 healthy controls matched group-wise with the patients on age and education (age: patients: $64.4 \pm 13.77$ years; controls: $67.5 \pm 6.0$ years; education: patients: $14.0 \pm 2.36$ years; controls: $14.7 \pm 2.4$ years). Patients with single, focal frontal lesions were identified from computed tomography and magnetic resonance imaging scans by a neurologist. Patients with lesions extending significantly into non-frontal regions were excluded. Detailed patient characteristics can be found in Baldo et al. (2002). Averaging across the five free recall trials, all three indices significantly separated patients from controls, with the $LBC_{sem}$ accounting for the largest amount of variance between the two groups (Original CVLT Semantic Clustering Index: $\omega^2 = .24$, $p = .025$; ARC $\omega^2 = .23$, $p = .027$; and $LBC_{sem}$, $\omega^2 = .29$, $p = .017$). Large sample studies are needed to determine whether the $LBC_{sem}$ is more sensitive to frontal lobe lesions, in particular, and other brain dysfunction, in general, than other clustering indices. However, these preliminary results hold out the promise that the $LBC_{sem}$ might be more sensitive than recall based indices.

## The Original CVLT Serial Clustering Index

As mentioned previously, the serial clustering index used by the original CVLT is calculated using a prediction equation. However, this equation does not adequately adjust for

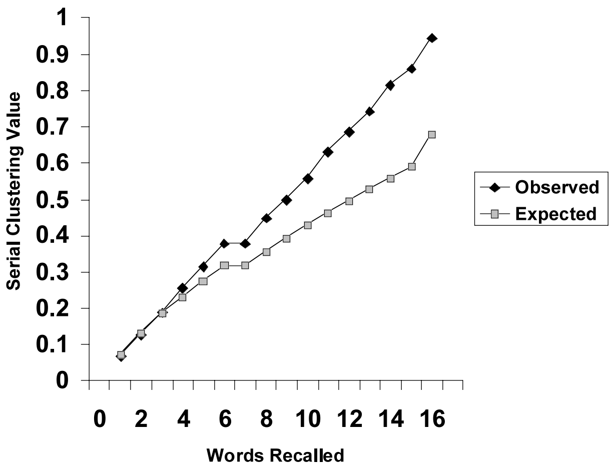


**Fig. 2.** Expected and observed serial clustering values calculated from simulations of random recall (original CLVT Index).

variations in total recall when serial clustering is random. As can be seen from Figure 1, the serial clustering index varies with the number of words recalled when serial clustering is random (this problem was also noted by Schmidt, 1997). Since no one serial clustering value marks random clustering, the serial clustering index should not be interpreted as demonstrating the amount of serial clustering above chance expectation. Figure 2 shows that the original CLVT Serial Clustering Index fails to adjust for chance clustering because the expected clustering value under-adjusts with greater recall. Given that the ratio of observed clustering to expected has different meanings for the CVLT Serial and Semantic Clustering indices, direct comparisons between amounts of serial clustering and amounts of semantic clustering are difficult to interpret.

## List-Based Serial Clustering Index

Table 2 compares LBC serial clustering index ($LBC_{ser}$) with the original CVLT's serial clustering index. As can be seen, the CVLT's serial clustering measure and $LBC_{ser}$ are similar in that they award more points for an increased number of observed serial clusters. However, $LBC_{ser}$ is not influenced by random recall. Table 3 demonstrates that the expected calculations for serial LBC based on Equation 7 are within rounding error of our Monte Carlo simulation.[4] As noted above, expected values for the original CVLT's serial clustering index are not independent of number of words recalled, even when recall is random (see Figure 1). In comparison with the original CVLT's serial clustering scores, $LBC_{ser}$ scores can be more clearly interpreted as the amount to which an individual's serial clustering exceeds what would be expected if that individual were randomly recalling words from the CVLT list without replacement. Additionally, if researchers are interested in the amount of bidirectional serial clustering (where a recall string including words from Position 2 then Position 1 would be counted as an observed cluster), then the expected value can be calculated by multiplying equation 7 by 2.

## CONCLUSIONS

To calculate an expected clustering value, the cluster space, that is, the number of possible clustering outcomes, must be clearly defined. Recall-based indices assume the number of words and categories recalled define the cluster space. By implication, any category clustering of words not recalled is unimportant. For recall-based indices it is not possible to determine how successfully an individual used clustering to master the individual list. List-based measures assume that the cluster space should be defined by the number of words and categories present in the original list. Such indices directly measure the joint occurrence of clustering and list mastery.

The measurement assumptions underlying recall-based and list-based cluster indices have different implications for theories relating organization to recall. The recall-based assumption that only categories and words present in the recall protocol are important to a clustering measure is compatible with a strong all or none account of category learning: The absence of a category in the recall data implies that learning involving that category did not occur. As mentioned in the historical section above, the assumptions

[4] Because smaller values were calculated, more precision was required to match the serial clustering values than to match the semantic clustering values. Consequently, a larger number of simulations was used to match our expected formula with simulation data (100,000 instead of 10,000).

**Table 2.** Comparing the original CVLT Serial Clustering Index with $LBC_{ser}$

| Example | Recall order | Words recalled | Observed clusters | Original CVLT Serial Ratio | $LBC_{ser}$ |
|---|---|---|---|---|---|
| 1 | 1, 2, 3, 4 | 4 | 3 | 16.32 | 2.81 |
| 2 | 1, 2, 5, 6 | 4 | 2 | 10.87 | 1.81 |
| 3 | 1, 2, 3, 4, 7, 10, 13, 14 | 8 | 3 | 8.45 | 2.56 |
| 4 | 1, 2, 3, 4, 5, 6, 7, 10, 12, 11, 14, 8, 13 | 14 | 6 | 10.75 | 5.19 |

*Note*. Numbers in the *recall order* column represent the serial position of a word on List A of the CVLT.

**Table 3.** Comparison of Monte Carlo based and equation based estimates of chance clustering

| Number of words recalled | Estimated chance semantic clustering: Monte Carlo | Estimated chance semantic clustering: Equation 5 | Estimated chance serial clustering: Monte Carlo | Estimated chance serial clustering: Equation 7 |
|---|---|---|---|---|
| 2 | 0.206 | 0.200 | 0.063 | 0.063 |
| 3 | 0.402 | 0.400 | 0.125 | 0.125 |
| 4 | 0.605 | 0.600 | 0.187 | 0.188 |
| 5 | 0.793 | 0.800 | 0.249 | 0.250 |
| 6 | 0.995 | 1.000 | 0.311 | 0.313 |
| 7 | 1.198 | 1.200 | 0.375 | 0.375 |
| 8 | 1.379 | 1.400 | 0.433 | 0.438 |
| 9 | 1.603 | 1.600 | 0.496 | 0.500 |
| 10 | 1.812 | 1.800 | 0.560 | 0.563 |
| 11 | 2.005 | 2.000 | 0.622 | 0.625 |
| 12 | 2.197 | 2.200 | 0.683 | 0.688 |
| 13 | 2.397 | 2.400 | 0.747 | 0.750 |
| 14 | 2.607 | 2.600 | 0.812 | 0.813 |
| 15 | 2.826 | 2.800 | 0.870 | 0.875 |
| 16 | 3.011 | 3.000 | 0.938 | 0.938 |

of recall-based measures are also compatible with the view that factors other than organization determine which words are recalled, with organizational processes occurring after list words are retrieved. However, the assumption that items from multiple categories are recalled silently and then overtly reported in a new and structured order is nearly impossible to reconcile with the timing of recall of categorized word lists (Wixted & Rohrer, 1994). Rather, people appear to use category names as retrieval cues, switching from one to another as recall progresses, but recalling words within a category as they come to mind.

Contrary to recall-based indices of clustering, list-based indices do not assume that only the words and categories recalled are relevant to measurement of organization. List-based indices are more compatible than recall-based indices with incremental theories of category learning. For list-based measures, the incremental mastery of a list is reflected in a graded increase of the list-based measure. Graded measurement is possible because list structure serves as an external standard that the recall protocol approximates in varying degrees.

As Shuell (1975) observed when considering the recall-based/list-based distinction, there does not seem to be a single best clustering measure. However, given the question of interest to the authors of the CVLT (i.e., the relationship between organization in encoding and recall), list-based measurements are preferable *because they allow for the possibility of such a relationship*. As a result, in the CVLT–II the recall-based clustering measures have been replaced with the list-based clustering measures that we have described in this paper.

Because the choice of a clustering measure is intimately tied to assumptions about memory processing, an alternative to using combinatorial indices to measure clustering, such as ARC or LBC, is to develop a parametric model of task performance where at least one parameter represents the process of category clustering. Several memory theorists have pursued this approach (Raaijmakers & Shiffrin, 1980; Riefer, 1982; Robertson, 1995). For models that use a single parameter to represent clustering, investigators could use the parameter value or some statistically desirable transformation of the parameter to measure clustering. As Colle (1972) observed, models that rely on multiple parameters to account for clustering call into question the use of a single summary index to measure clustering, as is the current practice. Whether a parametric model uses one or several parameters to account for clustering, such models provide the integration of theory and measurement missing in currently used combinatorial indices of category and serial clustering. Further progress in the measurement of the relationship between organization and recall is likely to depend on the development of memory theories that can be realized as parametric models.

## ACKNOWLEDGMENTS

Support for this research was provided by VA Merit Grants to Gregory G. Brown (VASDHC 304) and Dean C. Delis (WASDHS 316).

## REFERENCES


Baldo, J.V., Delis, D., Kramer, J., & Shimamura, A.P. (2002). Memory performance on the California Verbal Learning Test II: Findings from patients with focal frontal lesions. *Journal of the International Neuropsychological Society*, *8* (in press).

Bousfield, W.A. (1953). The occurrence of clustering in the recall of randomly arranged associates. *Journal of General Psychology*, *49*, 229–240.

Bousfield, A.K. & Bousfield, W.A. (1966). Measurement of clustering and of sequential constancies in repeated free recall. *Psychological Reports*, *19*, 935–942.

Bousfield, W.A. & Cohen, B.H. (1952). *The effects of reinforcement of the occurrence of clustering in the recall of randomly arranged associates*. Technical Report No. 1 under contract Nonr-631 (00) between the Office of Naval Research and the University of Connecticut.

Bousfield, W.A. & Cohen, B.H. (1955). The occurrence of clustering in the recall of randomly arranged words of different frequencies of usage. *Journal of General Psychology*, *52*, 83–95.

Bousfield, W.A. & Puff, C.R. (1964). Clustering as a function of response dominance. *Journal of Experimental Psychology*, *67*, 76–79.

Bousfield, W.A., Puff, C.R., & Cowan, T.M. (1964). The development of constancies in sequential organization during repeated free recall. *Journal of Verbal Learning and Verbal Behavior*, *3*, 489–495.

Buschke, H. (1975). Short-term retention, learning, and retrieval from long-term memory. In D. Deutsch & J.A. Deutsch (Eds.), *Short-term memory* (pp. 73–106). New York: Academic Press.

Colle, H.A. (1972). The reification of clustering. *Journal of Verbal Learning and Verbal Behavior*, *11*, 624–633.

Craik, F.I.M. (1981). Encoding and retrieval effects in human memory: A partial review. In J. Long & A. Baddeley (Eds.), *Attention and performance, IX* (pp. 383–402). Hillsdale, NJ: Lawrence Earlbaum Associates.

Crosson, B., Novack, T.A., Trenerry, M.R., & Craig, P.L. (1989). Differentiation of verbal memory deficits in blunt head injury using the recognition trial of the California Verbal Learning Test: An exploratory study. *Clinical Neuropsychologist*, *3*, 29–44.

Delis, D.C., Kramer, J.H., Kaplan, E., & Ober, B.A. (1987). *California Verbal Learning Test*. San Antonio, TX: The Psychological Corporation.

Frender, P. & Doubilet, P. (1974). More on measures of category clustering in free recall—although probably not the last word. *Psychological Bulletin*, *81*, 64–66.

Fridlund, A.J. & Delis, D.C. (1987). *CVLT Research edition administration and scoring software* [Computer software]. New York: The Psychological Corporation.

Hilgard, E.R. (1948). *Theories of learning*. New York: Appleton-Century-Crofts, Inc.

James, W. (1890). *The principles of psychology*. New York: Holt, Rinehart & Winston.

Mandler, G. (1967). Organization and memory. In K.W. Spence & J.T. Spence (Eds.), *The Psychology of Learning and Motivation*, *1* (pp. 327–372). New York: Academic Press.

Massman, P.J., Delis, D.C., Butters, N., Levin, B.E., & Salmon, D.P. (1990). Are all subcortical dementias alike? Verbal learning and memory in Parkinson's and Huntington's disease patients. *Journal of Clinical and Experimental Neuropsychology*, *12*, 729–744.

Miller, G.A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Murphy, M.D. (1979). Measurement of category clustering in free recall. In C.R. Puff (Ed.), *Memory organization and structure* (pp. 51–83). New York: Academic Press.

Pellegrino, J.W. & Hubert, L.J. (1982). The analysis of organization and structure in free recall. In C.R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 129–172). New York: Academic Press.

Roenker, D.L., Thompson, C.P., & Brown, S.C. (1971). Comparison of measures for the estimation of clustering in free recall. *Psychological Bulletin*, *76*, 45–48.

Raaijmakers, J.G. & Shiffrin, R.M. (1980). Search of associative memory. *Psychological Review*, *88*, 93–134.

Riefer, D.M. (1982). The advantages of mathematical modeling over traditional methods in the analysis of category clustering. *Journal of Mathematical Psychology*, *26*, 97–123.

Robertson, C. (1995). Modeling recall: Clustering and order effects. *British Journal of Mathematical and Statistical Psychology*, *48*, 29–50.

Schmidt, M. (1997). Some cautions on interpreting qualitative indices for word-list learning tests. *Clinical Neuropsychologist*, *11*, 81–86.

Shuell, T.J. (1969). Clustering and organization in free recall. *Psychological Bulletin*, *72*, 353–374.

Shuell, T.J. (1975). On sense and nonsense in measuring organization in free recall: Oops, pardon me, my assumptions are showing. *Psychological Bulletin*, *82*, 720–724.

Tulving, E. (1968). Theoretical issues in free recall. In T.R. Dixon & D.L. Horton (Eds.), *Verbal behavior and general behavior theory* (pp. 2–36). Englewood Cliffs, NJ: Prentice-Hall, Inc.

Underwood, B.J. & Schulz, R.W. (1960). *Meaningfulness and verbal learning*. Philadelphia: Lippincott.

Van Spaendonck, K.P.M., Berger, H.J.C., Horstink, M.W.I.M., Borm, F., & Cools, A.R. (1996). Memory performance under varying cueing conditions in patients with Parkinson's disease. *Neuropsychologia*, *34*, 1159–1164.

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


## Appendix A

### DERIVATION OF EXPECTED SEMANTIC CLUSTERING FORMULA

As indicated earlier, Frender and Doubilet's (1974) expected value formula is:

$$\mathrm{EXP}_{\mathrm{sem}} \text{ for a given trial} = \frac{[(r-1)(m-1)]}{N_{\mathrm{L}}-1} \tag{4}$$

Where $r$ = the number of correct words recalled on that trial, $m$ is the number of members of each semantic category on the original list (category size is assumed to be equal across categories), and $N_{\mathrm{L}}$ = the total number of words on the original list.

To illustrate why this formula yields the expected number of semantic clusters, we first compute the probability that the second item recalled falls into the same category as the first (i.e., we compute the probability that the second item yields a semantic cluster). After the first item has been recalled, there are $N_{\mathrm{L}} - 1$ remaining items that could be recalled. Of these $N_{\mathrm{L}} - 1$ items, $m - 1$ of them fall into the same category as the first recalled item. Thus, the probability that a random responder's second response will consist of an item from the same category as the first is simply $(m - 1)/(N_{\mathrm{L}} - 1)$. For the CVLT, this would be $(4 - 1)/(16 - 1)$, or 1/5.

For a random responder, the probability of a cluster appearing in the second output position is exactly the same as the probability of a cluster appearing in any other output position (save for the first, of course). For example, imagine a hypothetical response sequence consisting of $a, a, b, c, d$ in which a semantic cluster appears in output position 2. The probability of obtaining an item from category $a$ in the first output position by random chance on the CVLT is 4 (the number of items from category $a$ on the list) divided by 16 (the total number of items on the list). The probability of obtaining an item from category $a$ in the second output position by random chance is 3 (the number of not-yet-recalled items from category *a* on the list) divided by 15 (the total number of not-yet-recalled items on the list). Continuing in this way, the probability of obtaining the exact output sequence *a, a, b, c, d* is (4/16)(3/15)(4/14)(4/13)(4/12) = .001465. The probability that the *a, a* cluster would instead have appeared in output position 3 (*b, a, a, c, d*) by random chance is (4/16)(4/15)(3/14)(4/13)(4/12) = .001465.

The same value is obtained for the probability of the *a,a* cluster appearing in output positions 4 and 5, and the same exercise can be repeated for any possible output sequence. The critical point is that the probability of obtaining a cluster in output position 2 by random chance is the same as the probability of obtaining a cluster in any of the subsequent output positions by random chance. As indicated above, the probability of obtaining a cluster in position 2 by random chance is $(m - 1)/(N_L - 1)$. By induction, this is also the probability of obtaining a cluster in all of the subsequent output positions by random chance.

The expected number of clusters for a given recall protocol is simply the number of possible opportunities for a cluster to be observed (which is $n_j - 1$) multiplied by the probability of observing a cluster by random chance per opportunity, which is $(m - 1)/(N_L - 1)$. Thus, the expected number of clusters for a random responder is given by the formula supplied by Frender and Doubilet (1974).

## Appendix B

### DERIVATION OF EXPECTED SERIAL CLUSTERING FORMULA

The logic underlying the serial clustering formula is much like that of the semantic clustering formula given in Appendix A. We first compute the probability of obtaining a serial cluster in output position 2 and then rely on the fact that this is also the probability of observing a serial cluster in all subsequent output positions in order to derive an expected value.

The probability of observing a serial cluster in output position 2 depends on whether the item recalled in position 1 was the last item on the list or not. If the item recalled in output position 1 was the last item on the list, there is no chance of observing a serial cluster in output position 2. Thus, the probability of observing a serial cluster in position 2 is the probability that the first item recalled was *not* the last item in the list, which is $(N_L - 1)/N_L$, times the probability that the second item recalled is the one that followed the first in the original list. That probability is 1 (the number of candidate items) divided by $N_L - 1$ (the number of not-yet-recalled items). Thus, the probability of obtaining a serial cluster in output position 2 is $(N_L - 1)/N_L \times 1/(N_L - 1)$, which is simply $1/N_L$.

The probability of obtaining a serial cluster in output position 2 by random chance is the same as the probability of obtaining a serial cluster in the subsequent output positions by random chance. For example, consider the output sequence 10, 11, 4, 16, 9, which contains a serial cluster in output position 2. Assuming sampling without replacement, the probability of obtaining this exact sequence by random chance is (1/16)(1/15)(1/14)(1/13)(1/12), which is exactly the same as the probability of observing the output sequence 4, 10, 11, 16, 9 (serial cluster in output position 3) by random chance. Indeed, the probability is the same no matter where the serial cluster 10, 11 appears in the output sequence. The same kind of reasoning applies to any example. Thus, again by induction, one can conclude that the probability of observing a serial cluster by random chance in output position 2 is the same as the probability of observing a serial cluster in the subsequent output positions as well.

The expected number of serial clusters is simply the number of possible opportunities for a cluster to be observed (which is $n_j - 1$) multiplied by the probability of observing a cluster by random chance per opportunity, which is $1/N$. Thus, a slight modification of the formula supplied by Frender and Doubilet (1974) allows for serial clustering to be measured in a similar manner to semantic clustering.