Barriers to Repeated Assessment of Verbal Learning and Memory: A Comparison of International Shopping List Task and Rey Auditory Verbal Learning Test on Build-Up of Proactive Interference

S. Rahimi-Golkhandan1, P. Maruff1,2, D. Darby2, P. Wilson1,*

1Faculty of Arts and Science, School of Psychology, Australian Catholic University, Melbourne, Victoria, Australia
2CogState Ltd, Melbourne, Victoria, Australia

*Corresponding author at: Faculty of Arts and Science, School of Psychology, Australian Catholic University, 115 Victoria Pde., Fitzroy 3065, Australia.
Tel.: +61 3 9953-3445. E-mail address: peterh.wilson@acu.edu.au (P. Wilson).

Accepted 23 July 2012

Abstract

Proactive interference (PI) that remains unidentified can confound the assessment of verbal learning, particularly when its effects vary from one population to another. The International Shopping List Task (ISLT) is a new measure that provides multiple forms that can be equated for linguistic factors across cultural groups. The aim of this study was to examine the build-up of PI on two measures of verbal learning—a traditional test of list learning (Rey Auditory Verbal Learning Test, RAVLT) and the ISLT. The sample consisted of 61 healthy adults aged 18–40. Each test had three parallel forms, each recalled three times. Results showed that repeated administration of the ISLT did not result in significant PI effects, unlike the RAVLT. Although these PI effects, observed during short retest intervals, may not be as robust under normal clinical administrations of the tests, the results suggest that the choice of the verbal learning test should be guided by the knowledge of PI effects and the susceptibility of particular patient groups to this effect.

Keywords: Learning and memory; Assessment; Mild cognitive impairment; Test construction

Introduction

Neuropsychological (NP) tests of verbal memory are commonly administered repeatedly to the same individuals over relatively short retest intervals. For example, in clinical drug trials, verbal list-learning tests may be repeated at hourly intervals, whereas in the assessment of post-operative cognitive decline or recovery from concussion, these same tests may be given days or weeks apart (Lewis, Maruff, & Silbert, 2005). There is growing recognition that performance arising from the repeated application of memory tests can reflect cognitive processes other than memory. For example, repeated administration of the same list-learning test on separate assessments at short retest intervals allows one to develop a strong memory of the words themselves or to develop strategies to maximize recall, although these practice effects can be reduced through the use of parallel forms. However, the requirement to remember different word lists on repeated assessments can give rise to another form of cognitive change, where the words learned on one assessment will interfere with the recall of different words learned during subsequent assessments, termed proactive interference (PI). This PI manifests as decrease in accuracy of recall across repeated assessments. PI is often defined operationally as reduced total recall on subsequent lists compared with the original, and intrusion of words from earlier lists into the recall of a subsequent list. In instances where parallel forms of a test are administered serially, each more than once, PI is particularly salient on the first trial of a new list (Lustig, May, & Hasher, 2001).

Knowing about differences in susceptibility to PI across tasks may provide some important diagnostic/cognitive information for researchers and clinicians, alike. In other words, if a specific clinical group always shows higher/lower vulnerability to PI than do healthy controls, then PI may be used as a marker for identifying the disorder, or indeed, as a clinical characteristic.
However, structural differences between verbal learning tests such as stimulus type (Shulman, 1971), retrieval demands, and lists length (Ebert & Anderson, 2009) along with some methodological issues like the use of different indices of PI (Estevez-Gonzalez, Kulisevsky, Boltes, Otermin, & Garcia-Sanchez, 2003) influence the sensitivity of a test to PI and contribute to equivocal reports of PI effects in the literature (Belleville et al., 1992; Cushman, Como, Booth, & Caine, 1988; Kave & Heinik, 2004; Multhaup, Balota, & Faust, 2003; Schrijnemaekers, De Jager, Hogervorst, & Budge, 2006).

The magnitude of PI is increased when original and new words learned belong to the same semantic categories suggesting that the occurrence of PI reflects mainly semantic strategies for encoding (Rouleau, Imbault, Laframboise, & Bedard, 2001). Semantic encoding generally enhances retention and recall of unrelated word lists due to the binding of disparate information (Craik & Lockhart, 1972). However, in the case of NP assessment, there is the potential for words in different lists to have similar semantic associations which may enhance PI on subsequent recall (Kareken, Moberg, & Gur, 1996). Hence, the higher the number of shared semantic categories between the lists, the more likely PI is. For verbal list learning tests, the grouping of words by semantic category varies between three [Hopkins Verbal Learning Test (HVLT); Brandt, 1991], four [California Verbal Learning Test-II (CVLT-II); Delis, Kaplan, Kramer, & Ober, 2000], or none [Rey Auditory Verbal Learning Test (RAVLT); Schmidt, 1996]. Therefore, HVLT and CVLT-II might be more likely to cause PI than the RAVLT on which the items do not belong to particular categories. However, the lack of easily identifiable categories on a list can lead to PI, too. Given the reduced opportunity for grouping items into semantic categories on the RAVLT, one might need to visualize the words as being connected within a story and then use that story for recall. These so-called vivid stories also lead to PI because there is the potential for words in different lists to have similar semantic associations, which may enhance PI effects on subsequent recall (Kareken et al., 1996).

In a series of recent studies, we have outlined the sensitivity and cross-cultural validity of a novel verbal learning test, known as the International Shopping List Task (ISLT; Lim et al., 2009; Lim, Pietrzak, Synder, Darby, & Maruff, 2012; Pietrzak et al., 2009; Thompson et al., 2011). The use of a universal theme (i.e., shopping list items) in the ISLT may actually operate to decrease the PI associated with its repeated administration. Given that the category membership of shopping-list items is readily identifiable and thematic, the ISLT may provide retrieval cues that restrict the number of recall candidates and limit confusion on recall, thereby minimizing PI. Using familiar words also facilitates learning and recall without the need to form complex or novel semantic relationships (Kee & Helfend, 1977); this reduces the severity of PI. In contrast, recall of unrelated and/or unfamiliar nouns, as in the RAVLT, is more difficult because words cannot easily be grouped into specific categories (Groth-Marnat, 2003).

The effect of PI on the ISLT could be determined by comparing performance across repeated administrations to another validated list-learning test that does not use semantic categories at all, the RAVLT. The purpose of this study was to compare the build-up of PI between the ISLT and the RAVLT under conditions of repeated testing. We predicted that the repeated administration of the ISLT, at brief test–retest intervals, would result in significantly less PI than that of the RAVLT.

**Method**

**Participants**

The sample consisted of 61 (50 women) adults aged between 18 and 40 ($M = 21.5, SD = 3.6$). Participation was voluntary; however, given the type of assessments, participants needed to designate English as their primary language and have no current or past history of neurological or psychiatric disorders, substance, or alcohol dependence.

**Materials**

*International Shopping List Task.* The ISLT is a 16-word, three-trial verbal list-learning test. The lists consist of shopping items selected to be language and regionally commonplace. The presentation of lists and the recording of responses are controlled by a laptop computer. Each list is randomly generated from a pool of 128 words. A different set of 16 words is chosen when a new list is required, with eight different lists without overlap. A trial finishes as soon as no further recall is possible or 1 min has elapsed. A strong correlation of 0.84 between the ISLT and the HVLT established the criterion validity of ISLT. The ISLT has also demonstrated good test–retest reliability (Lim et al., 2009; Pietrzak et al., 2009).

*Rey Auditory Verbal Learning Test.* The RAVLT is a 15-word, five-trial, verbal list-learning test (Schmidt, 1996). The lists consist of widely used words in English that are short and concrete (high imagery), in that all are nouns for different objects. The RAVLT often includes five recall trials; however, only three recall trials were used here to be equivalent to the number ISLT trials. The RAVLT has three standard lists used as parallel forms. Test–retest reliability and the criterion
validity of the RAVLT are moderate (0.55 and 0.50–0.65, respectively) (Groth-Marnat, 2003; Macartney-Filgate & Vriezen, 1988). Although both tests also assess delayed recall, this study used the number of correctly recalled words and the number of intrusion errors on each trial as the primary outcome measures.

Procedure

Following approval from the RMIT Human Research Ethics Committee, undergraduate students at RMIT University were invited to participate. All participants gave their informed consent before data collection. Both tests were administered to the sample. However, in order to minimize the intrusion of items from one test to another, each test was administered in a separate session, separated by 1 week. The order of administration was counter-balanced over participants. Both tests had three parallel forms, and each form had three recall trials, with no interval between the trials of any one list. In order to control for differences in list length and facilitate the comparison of PI effects on the tests, an equal number of items (15 words) were used on each list. Following the presentation of a list (at the rate of one word per second), participants had 2 min to write down as many words as they could remember. Separate answer sheets were used for each trial, with earlier sheets removed from sight. Once the 2-min response period for the third trial of a list was over, participants were given a 5-min break before the experimenter started reading the next list.

Results

The mean total recall scores for each test and associated item lists are presented in Table 1. The observed difference between lists A and C on the RAVLT was more than two items, on average, but negligible for the ISLT. A two (test: ISLT and RAVLT) by three (list: A, B, and C) repeated-measures ANOVA on the average total recall found a significant interaction between test and list, Wilks’ $\Lambda = 0.90$, $F(2, 58) = 3.13$, $p = .05$, multivariate $\eta^2 = 0.10$. Tests of simple main effects revealed no significant difference between the ISLT lists, $p > .05$. However, the effect of list was significant for the RAVLT: Differences were shown between lists A and B ($p = .008$, $d = 0.29$) as well as lists A and C ($p = .001$, $d = 0.43$), but not between B and C ($p = .66$, $d = 0.14$). The effect size for the comparison between the first and last list on the RAVLT was 2.3 times larger than that for the ISLT.

The average number of correctly recalled words on the first trial of lists A, B, and C for each test is shown in Table 2. Recall averages for the three lists were similar on the ISLT. In contrast, for the RAVLT, there was a tendency for reduced recall across trials with each new list. A two (test: ISLT and RAVLT) by three (list: A, B, and C) repeated-measures ANOVA failed to show a significant interaction between test and list, Wilks’ $\Lambda = 0.96$, $F(2, 58) = 1.10$, $p = .34$, multivariate $\eta^2 = 0.04$. An analysis of simple effects, however, showed a significant difference in recall between lists A and C of the RAVLT ($p = .027$, $d = 0.37$). The difference between lists A and B of the RAVLT also approached significance ($p = .06$, $d = 0.28$). In contrast, the average recall was not shown to differ across lists for ISLT ($ps = .50$). Finally, word intrusion from prior lists was compared between lists B and C for each test. Minimal intrusions were recorded on both tests, with no significant difference between lists B and C for either of the tests.

Discussion

This study compared the build-up of PI between the ISLT and the RAVLT under conditions of repeated testing. Results were shown to support the hypothesis that the repeated administration of the ISLT results in significantly less PI than that of the RAVLT. The basis and implications of these results are discussed below.

Table 1. Means and standard deviations of total recall on each list of a test

<table>
<thead>
<tr>
<th>Test</th>
<th>List</th>
<th>$M$</th>
<th>$SD$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ISLT</td>
<td>A</td>
<td>35.37</td>
<td>4.03</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>35.60</td>
<td>4.79</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>34.45</td>
<td>5.32</td>
</tr>
<tr>
<td>RAVLT</td>
<td>A</td>
<td>34.62</td>
<td>4.42</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>33.20</td>
<td>5.36</td>
</tr>
<tr>
<td></td>
<td>C</td>
<td>32.41</td>
<td>5.50</td>
</tr>
</tbody>
</table>

Notes: ISLT = International Shopping List Task; RAVLT = Rey Auditory Verbal Learning Test.
The analyses of correct recall supported the hypothesis by showing no significant reduction in the average total recall of lists B and C of the ISLT, unlike the RAVLT which showed significantly higher recall on its first list. Differences in recall on the first trial of each list also corroborated this finding. Observed differences between the tests in terms of PI effects was in line with previous research suggesting that the stimulus type predicts the level of PI (Multhaup et al., 2003; Shulman, 1971). We suggest that the relationship between the stimulus type and PI is influenced by the type of learning strategy.

The categorization of words into specific clusters facilitates recall because category membership forms a cue which helps focusing only on relevant words (Saint-Aubin & Poirier, 2005). In instances where category membership is readily identifiable, clustering of words is relatively easy and one does not need to rely heavily on deeper levels of processing. In the case of the ISLT, the ease of grouping items on any list into particular clusters (e.g., fruits and snacks) restricts the number of recall candidates, reduces confusion in recall, and, consequently, lowers PI. In contrast, categorical clustering is less beneficial when words have minimal shared characteristics (e.g. the RAVLT which includes lists of unrelated words). The lower potential for grouping words into particular categories increases the number of recall candidates and improves the likelihood of error on recall. Therefore, individuals have to rely more on deeper levels of processing and relate words to each other in a way other than simple clustering. For example, they might visualize the words as being connected within a quirky story and then use that story for recall. However, these techniques that are based on creating so-called vivid stories also lead to PI because words in different lists may have similar semantic associations, which may enhance PI effects on subsequent recall.

Verbal learning tests are supposed to measure the capacity to simultaneously store and process currently relevant verbal information (Lustig et al., 2001). However, unresolved PI across successive test trials may act as a serious confound to the repeated assessment of verbal memory, reducing the sensitivity of the test. For instance, we observed a reduction of more than two words on total recall of the three parallel lists from the RAVLT. Because a clinically meaningful change on list learning tests can be demonstrated by a two-item reduction in total recall (Lim et al., 2012), this effect of PI may be incorrectly interpreted as a deficit. For the ISLT, however, this error of interpretation is less likely, based on our results.

No significant difference was observed between the tests in terms of overall intrusion errors. Likewise, previous research suggests that intrusions from previous trials are very infrequent in immediate serial recall (Hamilton & Martin, 2007; Saint-Aubin & Poirier, 2005). Moreover, those few studies that used more than one index of PI often reported inconsistent findings (e.g., Belleville et al., 1992). Negligible intrusion errors, in this study, could be attributed to the age of the sample. Given that older adults often show more intra-individual variability in task performance (Murphy, West, Armilio, Craik, & Stuss, 2007), we recruited young adults in order to minimize within-subjects variability and maximize the likelihood of detecting PI across both tests. However, some studies have suggested that there is a negative relationship between the level of cognitive functioning and likelihood of showing intrusion errors in recall (Lustig et al., 2001; May, Hasher, & Kane, 1999). Therefore, in our sample of young healthy university students, with a likely high level of cognitive functioning, minimal intrusions are consistent with previous research. As such, analysis of intrusions may not be the most robust test of the PI hypothesis here.

PI as a Confounding Factor

Researchers and clinicians alike need to be aware that observed differences between tests on recall patterns and PI effects may confound the assessment of verbal memory—a consistent pattern of performance cannot be assumed across tests. Rather, certain tests may afford different types of information about verbal learning and afford quite specific markers of dysfunction. However, if verbal memory is assessed with a specific test, using a specific index of PI, then the level of PI may assist in discriminating between different populations with known vulnerability to the effect.
According to May and colleagues (1999), performance on measures of working memory span is partially determined by the presence of interference within the task. Hence, the assessment of verbal memory is obscured when repeated administration of a test results in high PI. Likewise, higher-order cognitive functions (e.g., reasoning ability) among the elderly are best predicted by their working memory performance under low PI conditions (Emery, Hale, & Myerson, 2008). In short, it seems that those verbal learning tests that cause minimal PI yield more valid information about higher cognition in older adults.

However, PI may not always be a confounding factor in the assessment of verbal learning and memory. To illustrate, individuals with mild cognitive impairment or those in the early stages of Alzheimer’s dementia (AD) (who can hold a reasonable amount of information in working memory) may be distinguished from healthy adults by their higher susceptibility to PI. As such, PI may be a useful marker of early memory impairment in these groups—the RAVLT may be indicated to fully capture their range of learning issues. In contrast, adults with more severe memory impairment are less likely to experience the intrusions we associate with PI; as such, tests with higher PI effects will be less informative for this group.

**Future Research and Conclusions**

Although a sample of young adults maximized the likelihood of detecting PI, recruiting university students with possibly higher cognitive functioning than the general population not only restricted the generalizability of findings, but also contributed to minimal intrusion errors, and partially masked the actual differences between the tests. In addition, we used a very short retest interval, to enable us to better demonstrate that structural differences between the tests lead to different PI effects. However, very short retest intervals may have restricted application to clinical settings where cognitive change is evaluated on the basis of weekly or monthly re-assessments. For instance, the same level of PI may not be assumed under normal clinical administrations of the RAVLT. Therefore, a comparison of PI on the ISLT and other well-established verbal learning tests at longer retest intervals is still needed. Other future work will explore vulnerability to PI in different pathological conditions (like AD or schizophrenia) or in adult aging (Kave & Heinik, 2004; Schrijnemaekers et al., 2006).

Taken together, our results suggest that clinicians must be careful making inferences about change in episodic memory when these are based on repeated application of verbal learning tests (e.g., the RAVLT) over short retest intervals, for example, when examining the acute effects of drugs, or situational factors such as fatigue. Moreover, the choice of verbal learning test should be guided by knowledge of PI effects and the susceptibility of particular patient groups to this effect. This knowledge will hopefully help reduce errors of clinical interpretation.

**Conflict of Interest**

P Maruff is an employee of the CogState Ltd, the company which developed the International Shopping List Task (ISLT).

**References**


