Serial Killer: Investigating receptive vocabulary acquisition using word cards

Todd Leroux

Abstract

Cognitive psychology has played an integral role in the field of human information processing, memory and indeed learning in many areas. For over a century, there has been scientific interest in massed and distributed learning; and what's more, as light has been shone more intensely, the intricacies of how to maximize the potential benefits of distributed learning have emerged. Today, the value of distributed learning is highly respected as is the voluminous evidence supporting it. As a result, in the discipline of second language acquisition, the use of word cards and spaced repetitions is well entrenched. The light shone on the finer details has also revealed the influence of serial learning on vocabulary acquisition using word cards. This study will attempt to shine more light on this issue and contribute to the discussion by presenting experimental evidence to the question: To what degree does serial learning impact receptive vocabulary acquisition when using word cards?

Background

The purpose of this study is to investigate and measure the incremental differences between word card study tactics in a receptive context. Research has suggested that serial learning of word cards negatively impacts vocabulary acquisition (Nation, 2001; Norris, Baddeley, & Page, 2004). In this situation, it is not vocabulary learning that takes place exclusively as it is confounded by remembering the meaning of a word partly based on its relation to neighbouring items.

Subjects

Two intact classes of 1ˢᵗ year university students in an international program in a private Japanese university are the subjects and define the group composition: Group 1 — Serial Learning, Group 2 — Not Serial Learning. The study will occur as part of their regular course work, so internal validity is strong in this respect, as the subjects are not alerted to the experimental nature of their activities. In these subjects' regular class work, they will engage in extensive writing and TOEFL classes in which reading, writing and vocabulary are major components. Subjects also learn economics and use 'for native English speaker' textbooks. Students in this program, on average, have motivation levels higher than other students in other programs and indeed other universities (based on personal observation), which is an important consideration when discussing external validity.

All students in this program undergo a Vocabulary Size Test (Nation, 2008) and are subsequently assigned a vocabulary list to study. Typically, word lists such as the GSL and AWL are provided to the students and they receive regular vocabulary testing throughout their tenure in the program. What's more, subjects will have at least one and perhaps more standardized scores offering indications of proficiency level. In fact, subjects are placed in their respective classes based on standardized scores. As a result, inferences of relative ability pre-exist this study and are to be included as a covariate to determine whether the groups are or are not significantly different from each other, statistically. This fact also adds to the validity of this study as subjects are not subjected to any testing effects caused by pre-testing. Though using intact classes reduces the internal validity of this study, it is preferred to randomization due to reasons of within-group contamination. There is no doubt in the researcher's mind that contamination would occur using different treatments in the same group/class. Once again, though not randomized, group sameness can be determined through pre-existing proficiency data.

Methodology

In this study, to veil the true purpose, subjects will be advised that part of their course will be to determine which study technique is preferable to them. The subjects are normally studying lists but will now be instructed in the use of word cards. Subjects will be advised that they will be interviewed at a later date to

determine if word cards seemed a worthwhile study technique. They will also be advised that their knowledge level of the words would be assessed and is part of their course grade (albeit minuscule). This adds to the external and ecological validity strength of this study.

Materials

As the subjects' level has already been ascertained, they will select words from the British National Corpus (BNC) as found in the Range Program (Nation, n.d.). One criterion for initial selection of these words to be studied is that they are drawn from a level that starts from at least two 1000 word levels above highest size tested. This is not to say that this list will ensure that the subjects will not know or have not seen these words, rather it is a starting point for the creation of a common word card list.

Nonsense words were not selected as an option for this research because many of these students will go on to achieve high levels of English proficiency and will likely encounter the words selected in this study in the future. However, based on the pilot testing, if the envisioned word list approach is not effective, then nonsense words will have to be used. It is recognized that nonsense words are ethical and would be appropriate for this study if created according to proven and established guidelines.

Whether nonsense words are better for this study than the initial method selected is a question open for debate. Firstly, there is no pre-test in this study anyway, which is one benefit of using nonsense words, and thus not applicable. Next, nonsense words are beneficial as students are not as able to study target words outside of class; however, if students do engage in outside study and cannot find the target word in a dictionary, then questions and possible suspicions as to the purpose of their activities come into question. It is recognized that outside study of target words may occur and is an issue affecting validity. However, due to the busy schedules of the subjects outside of class, it is an accepted risk. A follow-up qualitative interview would aid in determining if outside study occurred.

Words selected for the student base list are based on the following literature supported criteria:

    a) Non AWL word list words. As the subjects use a native English speaker textbook in their course work, the chances for meeting these

words outside of the study would be high.

b) Non economics related words as per the above rationale.

c) Words with unrelated forms or meanings. (Nation, 2000; Erten and Tekin, 2008).

d) Words that do not contain obvious word parts (un, dis, re) as they would assist in comprehension by perhaps providing clues.

e) Poly-syllabic words - between 3 and 4 syllables - are to be selected. (Baddeley & And, 1975)

f) Salience - phonetically salient and spelling should provide a good indication of pronunciation i.e. no surprises - as a stable pronunciation will assist in retention. (Nation, 2001)

After receiving their initial word lists, subjects in each of the two groups will be instructed to check mark the first 50 unknown words on a vocabulary list that they believe they have not seen before. These lists will be collected and a common list of 20 - 30 words will be created (See Pilot Testing). In other words, chosen words will be the same between subject and between groups, thus controlling for item difficulty variability as much as possible and increasing the internal validity of this study. Because this will be a receptive-focused study, subjects will be provided an L1 translation, which is to be reviewed by 3 native speakers of Japanese. Subjects will be instructed in the use of word cards, but only with information regarding creation and basic study approach. Subjects in both groups will be provided 20 - 30 blank word cards each and they will create their cards and follow the study instructions under the guidance of the researcher/teacher.

Pilot testing:

Pilot testing strengthens the internal validity of any study. In this proposed study, pilot testing is absolutely essential as novel treatments and measures are being introduced. As such, there are numerous areas of uncertainty with regards how they will play out.

1. Word lists. The initial word lists to provided to the subjects for their 50-word study selection is of interest. Desirable are identical lists. This would ensure comparable or at least consistency with regards item difficulty though it is recognized that there will be variability in this respect due to individual differences. Pilot with 3 - 5 students of the same or higher level as subjects. If piloting not promising, then nonsense words to be used.

2. Number of word cards to be studied. Piloting will aid in the understanding of

what is a suitable number of word cards for this study though between 20 to 30 is expected (Nation, 2001) as the words have a relatively high level of difficulty. Specifically, floor and ceiling effects are of concern. Further, the number of words cards studied will impact the measurement aspect of this study. Pilot with the same 3 - 5 students of the same or higher level as subjects as above.

Treatment

Subjects in each group will study their set of word cards. The importance of time on task and the consequential effect on outcomes is well established in the field of vocabulary study. As such, it will be strictly controlled, which increases the internal validity of this study. Final decisions regarding time will be made after pilot testing. It is believed that between 45 minutes to 1 hour will be required to complete the word card creation process during the first session of the study. This must be determined with reasonable accuracy as class time is precious and this study occurs within an ongoing program. Also, and related to the number of word cards to be studied is an appropriate length of study time for each session keeping in mind the basic principles underlying the study: increased spacing, reduced time on task in each session is ultimately to be determined by pilot testing and with reasonable precision. The following is a possible schedule:

Session 1: Creation of cards and study - 60 minutes.

Session 2: (2 day lag) Serial Learning both groups - 20 minutes.

Session 3: (7 day lag) Serial Learning Group 1; Not Serial Learning Group 2 - 15 minutes.

Session 4: (14 day lag) Serial Learning Group 1; Not Serial Learning Group 2 - 10 minutes.

Session 4: Survey to gather data on attitudes to technique (also veils main purpose of study)

Session 4: Immediate post-test - 1st - 50% of target words.

Session 5: Delayed post-test 14 days later - 2nd - 50% of target words.

Session 6: Delayed post-test 60 days later - 100% of target words.

NOTE   Lags listed refer to time period after the preceding session.

Word cards will be collected after each session and remain in the possession of the researcher. For Group 2 - Not Serial Learning, the cards will be shuffled for Session 3 and shuffled again for Session 4. The subjects will be told that their cards have been used by other students and to just study them as they find them.

Pilot testing

Pilot testing would have to be more extensive in this area due to the very nature of the treatment.
1. Time on task. Floor and ceiling effects come into play here.
2. Session schedule. Floor and ceiling effects also come into play here. Is there enough time for to allow for acquisition of a suitable number of words; is there too much? Is spacing appropriate?

Measures

Sensitive vocabulary tests will be administered to test for degrees of learning. The test will include recognition and recall components as well as an attitude-centered survey. The attitude survey is primarily in place to veil the true focus of this study. However, if it can be developed into a valid and reliable instrument, then it might yield very interesting and pedagogically important data in terms of any interaction effect that may exist and thus account for a considerable amount of individual variance (See Figure 1). After completing the attitude survey there will be an immediate post-test.

There are two novel aspects of the immediate post-test in this study. First, 50% of the target words will be tested and randomly selected from word cards. Second, the immediate post-test will be administered in two parts - I and II. Part I of the immediate post-test is a recognition test that is an adaptation from Meara's (1999) 'Yes/No' test, and Wesche and Paribakht (1996) self-reporting instrument. The

Strongly agree Agree Disagree Strongly disagree

I think this technique:
1. Uses my study time effectively.
2. Will help me learn vocabulary.
3. Is better than studying lists.
4. Was a fun way to learn.
5. Is a waste of time.
6. Took too much time to make cards.
7. I can use again.

Figure 1　Measurement component: Immediate post-test Part I - Attitude survey item extract

target words are to be selected from a row of distractors. For the distractors, the first three letters can be the same as the target word but no more. For example, if a target word was 'absolute', then the distractors could be 'abysmal' and 'absolve'. The subject would put an 'X' to indicate the target word recognized. (See Figure 2) This is tantamount to 'I have seen this word before.' (Wesche and Parikbaht, 1996) Pilot testing of distractors is necessary to assess the emergent patterns of error. Patterns of error would suggest the distractor as inappropriate and mandate its removal from the list. After completion of Part I, forms are to be collected, and Part II administered. Part II will contain the target words in the recognition test, which is why Part I is collected.

Target words: absolute, malicious
Put an 'X' next to the word that appeared on your word card.
__ abysmal _X_ absolute ___ absolve ___ abrogate ___ advocate

__ majestic ___ malevolent ___ magnolia ___ marksmanship _X_ malicious

Figure 2　Measurement component: Immediate post test Part I - Recognition test extract

Part II will be a novel but sensitive multiple-choice test. There are two novel points of the multiple-choice test. First, the correct choice will be the actual translation provided to the subjects in that choices offered will require only partial

knowledge to choose correctly. However, the distractors for that target word will contain translations for the other target words on the subjects' word cards as well (See Figure 3). This is to minimize the possibility that the recognition of the L1 translation is the impetus for answer selection as opposed to the association of meaning to the target word.

---

1. <u>absolute</u> means   a) total and complete
                                        b) desire to harm
                                        c) wonderful
                                        d) stressed

For the question above, I feel:
                                        a) no confidence in my answer
                                        b) some confidence in my answer
                                        c) certain about my answer

Note: Definitions for multiple-choice are in L1 on actual test.

Distractor b) is the correct meaning for malicious, which is another target word. Distractors are recycled throughout the recall test as there will only be between 5 and 10 items (likely) tested. This is avoid the situation where seeing the translation is the impetus for answer selection as opposed to word - meaning association.

Self-report is in L1.

Figure 3   Measurement component: Immediate post test Part II - multiple-choice Recall test extract and self-report

The second novelty of Part II of the immediate post-test is that it will also include a self-report component. This component will follow each multiple-choice question. It will consist of the following choices: 'no confidence in my answer', 'some confidence in my answer', 'certain my answer is correct'. (See Figure 4) This self-report will be in the L1. Its inclusion in post-test is really a measure of strength of knowledge and could potentially be included as a variable for analysis though how to include it in the measure remains to be determined. Analyses with or without this component are easily accomplished.

8. <u>pragmatic</u> means _____ (Use Japanese)

For the question above, I feel：

          a）   no confidence in my answer

          b）   some confidence in my answer

          c）   certain about my answer

Self-report is in L1.

Figure 4 Measurement component: Immediate post test Part II - Receptive translation and self report

Finally, in Part II of the immediate post-test, there will be a separate component to assess word knowledge strength. In this component, the target word will be provided and the subject will be required to provide an L1 translation. A self-report is also included in the question format. Interestingly, answers that differ from the L1 translation originally provided to the subjects for word card creation may indicate outside of study review and thus warrant a qualitative investigation (interview with subject).

Fourteen days after the immediate post-test, a delayed post-test will be administered. It will follow the exact format as the immediate post-test (minus the attitude survey), but will contain the other 50% (i.e., non-tested target words). This is because using the same target words as the immediate post-test provides an additional encounter with a target word and thus confounds results. Finally, another delayed post-test approximately 60 days after the first delayed post-test (the end of term) will be administered using the entire set of target words and will also include the attitude survey. The attitude survey will be given after the recognition and recall tests purely out of researcher interest.

Scoring for the immediate post-test would be straightforward, which suggests external validity. Regarding the recognition test in Part I, 1 point for a correct answer - no partial credit. Regarding Part II, the recall component multiple-choice questions will also receive 1 point for a correct answer, and a correct L1 translation will receive 2 points. No partial credit will be given for either score. These scores will be analyzed as separate components and then analyzed as a composite score after a Rasch analysis.

Internal and construct validity is a major concern with this measurement tool

simply because of its novelty. Positively, it offers 3 separate measures of vocabulary items; however, extensive pilot testing is essential. With this design, a larger N size for the pilot is likely required to improve validity. Obviously, revisions and possibly even omissions from the test will be required.

Pilot testing

1. Attitude survey. Qualitative assessment though Rasch and factor analysis in future after development over time.
2. Distractors. Needed for both Part I and II looking for common errors.
3. Results including self-report. Review of data yielded. Run through Rasch and SPSS.
4. Implementation. Multiple forms so logistics should be observed including subject reactions and focus.

Analysis

Data will be run through a Rasch analysis. This is to convert raw data into actual measures. In addition, checks for outliers and confirmation of unidimensionality will be undertaken. After the measures are produced, an ANCOVA (Analysis of Covariance) will be run through SPSS. Results will be analyzed for significant differences between mean scores for between groups. Due to the relatively small N size and items to be tested, effect sizes may not present themselves. In other words, the study may not have enough power to yield significant results. In closing, for future analysis other data could also be easily analyzed, such as differences based on gender, for example, to assess the existence of interaction effects.

Summation

What began as a simple question of a well-researched area, contemplating the design of this study led to many more questions. Though much thinking time, effort and literature review has been performed to keep the design simple, it has become intricate and complex. Efforts to account for every confounding variable and every factor reducing validity have perhaps detracted from the practicality of this study.
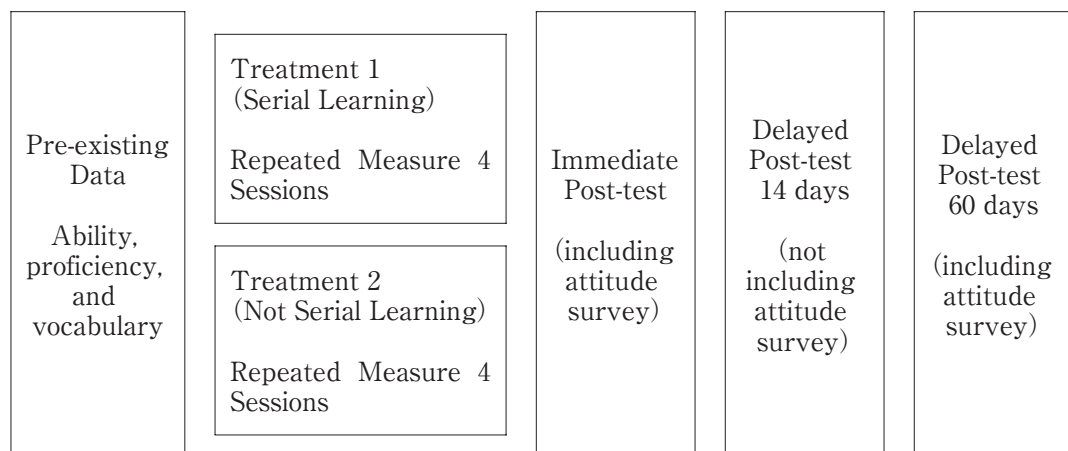
However, confidence in the justifications throughout is strong. Save for the measurement aspect of this study, which is the big unknown due to its novelty, results produced are likely to show what was expected: that serial learning negatively impacts receptive vocabulary acquisition. Further, if a value could be attached such as a percentage to this finding, then students might be more willing to adopt a specific learning strategy (cognitive in this case) and use it correctly.

Executive Summary

Research question: To what degree does serial learning impact receptive vocabulary acquisition when using word cards as a study technique?

Experimental Design:

Pre-existing Data Ability, proficiency, and vocabulary	Treatment 1 (Serial Learning) Repeated Measure 4 Sessions Treatment 2 (Not Serial Learning) Repeated Measure 4 Sessions	Immediate Post-test (including attitude survey)	Delayed Post-test 14 days (not including attitude survey)	Delayed Post-test 60 days (including attitude survey)

Pilot testing:
1. Materials
   - Initial word lists (Subjects select unknown words)
   - Number of word cards to be studied (20 - 40 of equal burden)
2. Treatment
   - Time on task requirements
   - Non-serial learning schedule
3. Measures
   - Attitude Survey
   - Distractors

- Results including self-report data
- Implementation

NOTE: Pilot testing to determine exact numbers/amounts/times listed below.

Subjects: 2 intact classes of 1st year (private) university students in Japan who are registered in an International Economics Program.
n = 40 (2 groups x 20 per group).
Group 1 - Serial Learning; and Group 2 Not Serial Learning

Materials: Subjects select 50 unknown words of equal burden from list.
Researchers select 20 - 30 common words for word card creation and include
L1 translation.
Subjects begin treatment.
Word cards kept by researcher.

Treatment: Session 1: creation of cards and study - 60 minutes.
Session 2: (2 day lag) Serial Learning both groups - 20 minutes.
Session 3: (7 day lag) Serial Learning Group 1; Not Serial Learning Group 2 - 15 minutes.
Session 4: (14 day lag) Serial Learning Group 1; Not Serial Learning Group 2 - 10 minutes.
Session 4: Survey to gather data on attitudes to technique (also veils main purpose of study)
Session 4: Immediate post-test - 1st - 50% of target words.
Session 5: Delayed post-test 14 days later - 2nd - 50% of target words.
Session 6: Delayed post-test 60 days later - all target words.

Measures: Pre-existing Vocabulary Size Test (Nation) scores, TOEFL scores
Measure I: Attitude survey (also used to veil main study purpose)
Measure II: Recognition test (adapted Meara, 1999)
Measure III: Level matching test (Nation, 1983) + self-report
Measure IV. Translation L1 (Receptive) + self-report

Independent variables: Serial learning, Non-serial learning

Rasch analysis (unidimensionality and outlier identification)

Dependent variable: Mean scores (ANCOVA - SPSS)

Validity

Internal

Aspect of the study	Application	Rating
Subjects	Able to be determined though not randomized.	Medium
Materials	Texts are same in both treatments. Words are same in both treatments. Pilot testing.	Strong
Treatments	Consistently applied to all subjects. Time on task the same. All subjects are equally familiar with the task. Surrounding conditions equal for all treatments. *Repetitions will vary as based on individual subject (not controlled). Outside of study review possible (not controlled). Pilot testing.	Medium
Measures	Measures are the same for both treatments. Measures are administered and scored the same for both treatments. *Novelty of measures is a concern, especially self-report scoring procedure.	Unknown

Ecological

Aspects of the study	Application	Rating
Subjects	Typical language learners.	Strong

Materials
  Texts
  Words
	Same source (British National Corpus).
	Typical words though difficult.
	Semantic relationship considered (similarities avoided).
	Decontextualized (as intended).
	Word length controlled.
	Part of speech controllable.
		Strong to medium

Treatment	Treatments are like normal learning activity as strategies are commonly addressed. Subjects not aware of experiment.	Strong
Measures	Immediate and Post-tests identical in format, but use different in target words (1st test - 50%, 2nd test - other 50%, 3rd test - 100%). 3 measures for each target word. Measures relevant for each word. *Self-report measure to be determined or excluded.	Strong
Pedagogical significance	N size and limited number of items to be tested is likely to produce 'Power coefficient' issues.	Medium to weak

References

Acheson, D., & MacDonald, M.（2009）. Verbal Working Memory and Language Production: Common Approaches to the Serial Ordering of Verbal Information. *Psychological Bulletin, 135*（1）, 50-68. Retrieved from ERIC database.

Baddeley, A., & And, O.（1975）. *Word Length and the Structure of Short-Term Memory*. Journal of Verbal Learning and Verbal Behavior. Retrieved from ERIC database.

Bloom, K., & Shuell, T.（1981）. Effects of Massed and Distributed Practice on the Learning and Retention of Second-Language Vocabulary. *Journal of Educational Research, 74*（4）, 245-48. Retrieved from ERIC database.

Bower, B.（1987）. Memory Boost from Spaced-Out Learning. *Science News, 131*（16）, 244. Retrieved from ERIC database.

Cepeda, N., Pashler, H., Vul, E., Wixted, J., & Rohrer, D.（2006）. Distributed Practice in Verbal Recall Tasks: A Review and Quantitative Synthesis. *Psychological Bulletin, 132*（3）, 354-380. Retrieved from ERIC database.

Erten, I., & Tekin, M.（2008）. Effects on Vocabulary Acquisition of Presenting New Words in Semantic Sets versus Semantically Unrelated Sets. *System: An International Journal of Educational Technology and Applied Linguistics, 36*（3）, 407-422. Retrieved from ERIC database.

Nakata, T.（2008）. English Vocabulary Learning with Word Lists, Word Cards and Computers: Implications from Cognitive Psychology Research for Optimal Spaced Learning. *ReCALL, 20*（1）, 3-20. Retrieved from ERIC database.

Nation, P.（2001）. Learning vocabulary in another language. Cambridge, England: Cambridge University Press.

Norris, D., Baddeley, A., & Page, M.（2004）. Retroactive Effects of Irrelevant Speech on Serial Recall From Short-Term Memory. *Journal of Experimental Psychology Learning Memory and Cognition, 30*（5）, 1093-1105. Retrieved from ERIC database.

Paribakht, T. S., & Wesche, M. B.（1993）. Reading comprehension and second language development in a comprehension-based ESL program. TESL Canada Journal 11（1）, 9-29.

Pimsleur, P.（1967）. *A MEMORY SCHEDULE*. Retrieved from ERIC database.

Read, J.（1993）. The development of a new measure of L2 vocabulary knowledge.

Language Testing, 10, 355-371.

Read, J. (2000). Assessing vocabulary knowledge and use. Cambridge, England: Cambridge University Press.

Vollmer, M., & And, O. (1989). The Effects of Two Types of Instruction on Simultaneous and Sequential Processing. Retrieved from ERIC database.

Webb, S. (2007). The Effects of Repetition on Vocabulary Knowledge. *Applied Linguistics, 28* (1), 46-65. Retrieved from ERIC database.

Willson, V. (1982). Maximizing Reliability in Multiple Choice Questions. *Educational and Psychological Measurement, 42* (1), 69-72. Retrieved from ERIC database.