LABOR AND GAIN, BUT THEN LABOR IN VAIN: DIMINISHING RETURNS OF REPEATED RETRIEVAL PRACTICE

A thesis submitted
To Kent State University in partial
Fulfillment of the requirements for the Degree of Master of Arts

By
Mary A. Pyc
May, 2008
Thesis written by
Mary A. Pyc
B.S., Keene State University, 2005
M.A., Kent State University, 2008

Approved by
Katherine A. Rawson, PhD  Advisor
Mary Ann Stevens, PhD  Chair, Department of Psychology
Jerry Feezel, PhD  Dean, College of Arts and Sciences

ii
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>LIST OF FIGURES</td>
<td>iv</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>v</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>vi</td>
</tr>
<tr>
<td>INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>INTRODUCTION EXPERIMENT 1</td>
<td>7</td>
</tr>
<tr>
<td>METHODS</td>
<td>8</td>
</tr>
<tr>
<td>Materials</td>
<td>8</td>
</tr>
<tr>
<td>Results</td>
<td>11</td>
</tr>
<tr>
<td>INTRODUCTION EXPERIMENT 2</td>
<td>15</td>
</tr>
<tr>
<td>METHODS</td>
<td>17</td>
</tr>
<tr>
<td>Materials</td>
<td>17</td>
</tr>
<tr>
<td>Results</td>
<td>20</td>
</tr>
<tr>
<td>GENERAL DISCUSSION</td>
<td>26</td>
</tr>
<tr>
<td>FUTURE DIRECTIONS</td>
<td>31</td>
</tr>
<tr>
<td>REFERENCES</td>
<td>36</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

FIGURES

1 Final test performance as a function of the number of times an item was correctly recalled during practice, interstimulus interval (ISI), and retention interval (RI). All standard errors of the mean were between 1.1 and 7.4.

2 Mean final test performance for the ISI 34 short RI group and the ISI 34 long RI group as a function of the number of times an item was correctly recalled during practice, and fits of the full and reduced models to those data.

3 Mean final test performance for the ISI 6 short RI group and the ISI 34 short RI group as a function of the number of times an item was correctly recalled during practice, and fit of the full model to those data.
LIST OF TABLES

TABLES

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Final test performance as a function of kind of response on first test-restudy trial for repeated items</td>
</tr>
<tr>
<td>2</td>
<td>Final test performance as a function of the number of times correctly recalled during practice</td>
</tr>
</tbody>
</table>
ACKNOWLEDGMENTS

This project would not have been possible without the assistance, guidance, and support of several individuals. Specifically, I thank Katherine Rawson for advising me on this project. I would also like to thank the members of my committee: John Dunlosky, Maria Zaragoza, and Dan Neal.
INTRODUCTION

Traditionally, testing has been used primarily as a means for instructors to assess how well students know material for a given course. More recently, however, it has been acknowledged that testing is not only beneficial as a means of assessment, but is also a powerful tool to aid students in learning of course materials that need to be recalled on examinations (e.g., Bahrick, Bahrick, Bahrick & Bahrick, 1993; Carpenter & DeLosh, 2005; Herrmann, Buschke, & Gall, 1987; and more recently, Roediger & Karpicke, 2006).

Simply stated, testing effects refer to the finding that higher levels of final test performance are attained when items are tested as opposed to restudied during practice (e.g., Carpenter, Pashler, & Vul, 2006; Chan, McDermott, & Roediger, 2006; McDaniel & Masson, 1985; Roediger & Karpicke, 2006), even when controlling for the amount of time spent studying or testing (e.g., Carrier & Pashler, 1992). This general pattern has been obtained predominantly with word lists and paired associates (e.g., Bahrick & Phelps, 1987; Cull, 2000); however, recently this same testing effect has emerged when more complex materials such as pictures (Roediger & Payne, 1982) and notes from college course lectures (e.g., Chan, McDermott, & Roediger, 2006) are utilized. Additionally, the benefits of testing during practice are found even when the final test assesses memory for related materials, not directly tested or studied during practice. For example, using complex materials (e.g., an article about toucans), Chan et al. (2006)
investigated whether testing effects extend to materials that are related to, but not directly tested during practice. All participants first read an article about the characteristics and habitat of the toucan. After reading the article, half of the participants answered short questions from the article (e.g., “where do toucans sleep at night?”), and the other half of the participants read short facts from the article (e.g., “toucans sleep in tree holes at night”). After a one day delay, all participants returned for a final test which assessed memory for information related to the facts or questions that they answered or studied on the previous day. Results demonstrated that final test performance for related information was enhanced for the test group, but not for the study group.

Although final test performance is higher when items are tested versus studied during practice, the combination of both testing and restudying items during practice leads to higher levels of final test performance than either testing or restudying alone (Carpenter et al., 2006; Cull, 2000; Karpicke & Roediger, 2007). For example, in a series of experiments, Cull (2000) utilized paired associates to assess final test performance as a function of practice schedule. More specifically, Cull (2000) compared final test performance levels for test-restudy, test-only, and study-only practice schedules of learning. As in other research showing a testing effect, test-only practice schedules outperformed study-only practice schedules. More important, results demonstrated that test-restudy practice schedules outperformed both test-only and study-only practice schedules. Thus, although there is a benefit for testing compared to studying during practice, the combination of both testing and restudying during practice leads to higher levels of final test performance than either alone.
Additionally, multiple test-restudy practice trials lead to higher levels of final test performance than a single test-restudy trial (provided that they are distributed across time and/or materials; e.g., Bahrick & Hall, 2005; Cull, 2000, Karpicke & Roediger, 2007; Pyc & Rawson, in press). For example, Pyc and Rawson (in press) investigated the effectiveness of test-restudy practice schedules using paired associates. Of interest here, Pyc and Rawson (in press) implemented schedules in which items received one test-restudy practice trial or three test-restudy practice trials after initial study. Results indicated higher levels of final test performance for items receiving three test-restudy practice trials compared to items receiving only one test-restudy practice trial.

Although previous research has shown that multiple distributed test-restudy trials can be effective for promoting memory, an outstanding question remains: How much is enough? That is, at what point can test-restudy practice with a given item be discontinued without compromising subsequent memory for that item? The primary goal of the two experiments presented herein was to address this question by investigating the relationship between amount of practice—and in particular, the number of times an item is correctly recalled during practice—and subsequent memory at final test. Exploring this question may have important implications for both application and theory. For example, consider a college student who is studying for final exams. Not only does each exam require memory for a large amount of material, but students often have 4 or 5 exams that they are studying for at the same time. Thus, students have a large amount of information that they need to know, but a limited amount of time in which to learn it. Additionally, being able to do well on exams becomes increasingly important as students
progress into high school and college where their grades determine whether they will be able to pursue higher educational degrees. Therefore, it is important for students to study materials enough so that they will be able to remember them later, but they must also be economical in the amount of time they continue studying material they know so that sufficient time will be available to study material that they have not yet learned well.

Regarding theoretical implications, the present paper will provide important new patterns of data that will help guide and constrain theories of testing effects. More specifically, the data can be used to further explore the implications of some proposed theoretical accounts of testing effects, such as the retrieval hypothesis (e.g., Glover, 1989) and the desirable difficulties account (e.g., Bjork, 1994). Thus, the primary goal of the two experiments in the current paper was to investigate the relationship between number of correct recalls during practice and subsequent memory performance at final test.

Pyc and Rawson (in press) recently reported results relevant to the question concerning the relationship between the number of correct recalls during practice and subsequent memory performance at final test. Of interest here, they compared two schedules of test-restudy practice using Swahili-English paired associates. In the fixed schedule, each item received three practice test-restudy trials (as in most previous research). In the dropout schedule, items were only practiced until they were correctly recalled once during practice. Only those items that were not correctly recalled on a previous practice test received further test-restudy trials. In two experiments, fixed and dropout schedules led to similar levels of performance on a final cued recall test.
administered after a short delay. Thus, a dropout schedule, in which items could only be correctly recalled one time during practice, led to similar levels of final test performance as a fixed schedule, in which items could be correctly recalled up to three times.

Recently, however, Karpicke and Roediger (2007) found different results than Pyc and Rawson (in press). Of interest here, Karpicke and Roediger’s (2007) study included a fixed schedule in which each item received eight test-restudy trials and a dropout schedule in which test and restudy were dropped after an item was correctly recalled once during practice. Results showed that performance on a final free recall test one week later was significantly worse in the dropout condition than in the fixed condition (21% versus 44%, respectively).

Of course, the methods used in these two studies differed from one another in several respects. Karpicke and Roediger (2007) utilized free recall of word lists, whereas Pyc and Rawson (in press) implemented cued recall of Swahili-English paired associates. Additionally, the fixed schedule in Karpicke and Roediger’s study included eight test-restudy trials, whereas Pyc and Rawson’s fixed schedule included three test-restudy trials. However, I suspected that the methodological difference most likely responsible for the diverging results is the retention interval. More specifically, Pyc and Rawson (in press) implemented a short retention interval (within session), whereas Karpicke and Roediger (2007) implemented a longer retention interval (one week, between session).

Thus, with respect to my main question concerning the relationship between the number of correct recalls during practice and subsequent memory performance, results are mixed. To further investigate this question, I conducted two experiments to examine
the relationship between the number of correct recalls during practice and subsequent memory performance for materials, with a primary interest in the extent to which the effect of the number of correct recalls during practice depends on retention interval. In the first experiment, I utilized the same basic paradigm as Pyc and Rawson (in press) but implemented a one week retention interval, to investigate whether retention interval was the variable responsible for the difference in results between Pyc and Rawson (in press) and Karpicke and Roediger (2007). To foreshadow, I found converging evidence with Karpicke and Roediger (2007), showing that discontinuing practice after one correct recall compromised performance at final test. In Experiment 2, I further explored the relationship between the number of correct recalls during practice and subsequent memory for those items on a final retention test, and the extent to which this relationship depended on retention interval.
INTRODUCTION EXPERIMENT 1

To investigate the relationship between the number of correct recalls during practice and subsequent memory performance, I implemented two schedules of learning: a fixed schedule and a dropout schedule. Both of these schedules had 23 intervening items between initial study and first practice test, as in Pyc and Rawson, Experiment 2 (in press). Importantly, in contrast to Pyc and Rawson (in press), in the present experiment I implemented a one week retention interval as in Karpicke and Roediger (2007).
Participants and design. Twenty-six undergraduate students enrolled in Introductory Psychology at Kent State University participated in return for course credit. Schedule of practice (fixed or dropout) was a within-participant variable.

Materials. Two lists of 24 Swahili-English translation word pairs were utilized with an equivalent range of item difficulty on each word list (based on norms reported by Nelson & Dunlosky, 1994). Assignment of list to practice schedule condition was counterbalanced across participants.

Procedure. All task instructions and items were presented via computer. Half of the participants completed the fixed schedule first followed by the dropout schedule, and the other half of the participants completed the dropout schedule first followed by the fixed schedule. For each practice schedule, each of the 24 items received an initial study trial followed by test-restudy practice for that list. On an initial study trial, the Swahili word was presented on the left side of the screen and the target English translation appeared on the right side of the screen for ten seconds. During test-restudy practice trials, the Swahili word was presented alone and participants had eight seconds to enter the English translation in a text box provided below the Swahili word. After eight seconds had elapsed, the response box was removed from the screen, and the Swahili and English words were presented together for a four second restudy opportunity.
In the fixed condition, each item was presented for three test-restudy practice trials, with an interstimulus interval (ISI) of 23 items between initial study and each subsequent test-restudy practice trial.

In the dropout condition, 23 items intervened between initial study and the first test-restudy practice trial. If the correct translation for the Swahili word was recalled, the item was dropped from further practice. If the translation for the Swahili word was not correctly recalled, the item was placed at the end of the list of to-be-learned items for another test-restudy practice trial. This process continued until either all of the items were correctly recalled once, or a participant reached the 72 maximum trial allowance (not including the initial study trials). The maximum allowance of test-restudy practice trials was set at 72 so that the amount of practice in the dropout condition did not exceed the 72 test-restudy practice trials allotted in the fixed condition, as in Pyc and Rawson (in press). This stopping rule affected 9 participants. On average, participants used 57.7 trials (SE = 2.8) in the dropout condition.

Upon completion of the practice phase of the experiment, participants were dismissed and reminded to return for their second session one week later. The final test was a participant paced cued-recall test.
RESULTS EXPERIMENT 1

The mean percentage of items correctly recalled at final test in the fixed and dropout conditions was 31.4% (SE = 4.5) and 14.3% (SE = 2.6), respectively. A paired-samples t-test revealed a statistically significant difference between fixed and dropout for overall levels of final test performance; \( t(25) = 4.66, p < .001 \). Results thus contrast those reported previously in Pyc and Rawson (in press), and provide converging evidence with Karpicke and Roediger’s (2007) results: One correct recall during practice was not enough to obtain the same level of performance as in a condition that afforded more than one correct recall, when the final test was administered after a one week delay.

To investigate why I found a different pattern of results in this study than in Pyc and Rawson (in press), I conducted conditional analyses to compare to those reported by Pyc and Rawson (in press). Specifically, as in Pyc and Rawson (in press), I computed final test performance as a function of the status of an item at first practice test (i.e., correctly or incorrectly recalled at first practice test). For ease of comparison, results of conditional analyses from both studies are presented in Table 1. In Pyc and Rawson (in press), items that were initially correctly recalled at first practice test were significantly more likely to be correctly recalled at final test in the fixed condition compared to the dropout condition. Similarly, in the present experiment items that were initially correctly recalled at first practice test were significantly more likely to be correctly recalled at final test in fixed compared to dropout, \( t(25) = 5.21, p < .001 \). However, the advantage for
initially correct items in the fixed condition over the dropout condition appears to be
greater after a week delay.

Table 1: Final Test Performance as a Function of Kind of Response on First Test-
Restudy Trial for Repeated Items, Experiments 1

<table>
<thead>
<tr>
<th></th>
<th># items correct on 1st test</th>
<th>Probability of recall on final test given correct on 1st test</th>
<th># of items incorrect on 1st test</th>
<th>Probability of recall on final test given incorrect on 1st test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pyc &amp; Rawson, Experiment 2 (in press)</td>
<td>5.6 (.94)</td>
<td>.85 (.06)</td>
<td>18.4 (.94)</td>
<td>.46 (.06)</td>
</tr>
<tr>
<td>Fixed</td>
<td>7.9 (.98)</td>
<td>.67 (.06)</td>
<td>16.1 (.98)</td>
<td>.66 (.06)</td>
</tr>
<tr>
<td>Dropout:</td>
<td>7.0 (1.0)</td>
<td>.59 (.07)</td>
<td>17.0 (1.0)</td>
<td>.24 (.04)</td>
</tr>
<tr>
<td>Experiment 1</td>
<td>7.9 (.93)</td>
<td>.16 (.04)</td>
<td>16.1 (.93)</td>
<td>.13 (.03)</td>
</tr>
</tbody>
</table>

Note: Values in parentheses are standard errors of the mean. Total number of items in
each group was 24.

Concerning items that were initially incorrectly recalled on the first practice test,
Pyc and Rawson (in press) found that they were significantly more likely to be correctly
recalled at final test in dropout compared to fixed. Interestingly, I did not find this same
pattern in the current experiment. Items that were initially incorrectly recalled at first
practice test were significantly less likely to be correctly recalled at final test in dropout
than in fixed; t(25) = 2.73, p = .01.
Why were initially incorrectly recalled items less likely to be recalled on the final test in the dropout condition versus the fixed condition? Additionally, for initially correct items, why was the difference in final test performance between fixed and dropout much larger here than in Pyc and Rawson (in press)? The answers may concern the total number of times items were correctly recalled during practice. First, consider the fate of initially correctly recalled items in the two conditions. With the dropout schedule, items that are initially correctly recalled during practice are dropped from test-restudy practice. In contrast, with the fixed schedule, items that are initially correctly recalled during practice have two more test-restudy practice trials. Thus, items in the fixed condition could be correctly recalled up to three times during practice. Whereas one correct recall during practice may be sufficient to retain information across a short delay, one correct recall is apparently not sufficient to retain information across a longer delay. Similarly, consider the fate of initially incorrectly recalled items in the two conditions. With the dropout schedule, items that are initially incorrectly recalled continue to receive test-restudy practice, but only until they are correctly recalled once. In contrast, with the fixed schedule, items that are initially incorrectly recalled receive two more test-restudy practice trials and thus may be correctly recalled up to two times during practice.

To further explore the extent to which differences between the fixed and dropout conditions in final test performance were due to differences in the number of times that items were correctly recalled during practice, I conducted a second set of conditional analyses. Specifically, I examined final test performance as a function of the number of
times an item was correctly recalled during test-restudy practice. Means for the second set of conditional analyses are presented in Table 2.

Table 2: Final Test Performance as a Function of the Number of Times Correctly Recalled During Practice

|                                          | Fixed          | Dropout        |
|                                          | M  | SE | M  | SE |
|------------------------------------------|----------------|----------------|
| FT performance for items correct 0 times during practice | .08 | .05 | .01 | .01 |
| For items that were initially incorrectly recalled: | .19 | .06 | .14 | .03 |
| For items that were initially correctly recalled: | .13 | .05 | .16 | .04 |
| FT performance for items correct 2 times during practice |                      |
| For items that were initially incorrectly recalled: | .44 | .05 |       |    |
| For items that were initially correctly recalled: | .42 | .20 |       |    |
| FT performance for items correct 3 times during practice | .45 | .09 |       |    |

Because I did not manipulate the number of times items were correctly recalled during practice, not surprisingly, not all participants contributed a value to each cell of the table. Therefore, I did not run inferential statistics on these data. Instead, I simply point the reader to Table 2 to see the general pattern. Of course, interpretations of these data should be made with some caution due to the differential number of participants contributing to each cell, as well as possible item difficulty effects (items correctly
recalled more times during practice were likely the normatively easier items).
Nonetheless, the overall pattern suggests that final test performance depends on the number of correct recalls during practice, regardless of the initial correct or incorrect status of items at first practice test or the amount of practice trials items received. For example, items that were correctly recalled once during practice obtained similar levels of final test performance regardless of whether they were initially correct or incorrect on the first test trial. Similarly, the total amount of practice did not appear to matter. For example, in the fixed condition, all items had three practice trials, but final test performance differed as a function of the number of times an item was correctly recalled during those practice trials. Consider also the finding that final test performance in the fixed and dropout conditions was similar for initially correct items that were only correctly recalled once during practice (.13 and .16), despite the fact that practice was discontinued in the dropout condition whereas in the fixed condition items had two additional practice trials.

Although the above comparisons suggest that correctly recalling an item more times during practice is beneficial for final test performance, is it always beneficial to recall items more times during practice? Interestingly, results of these analyses also suggest diminishing returns, given that there was not a difference in final test performance for those items that were correctly recalled two versus three times during practice. Experiment 2 was designed to more systematically explore these patterns.
INTRODUCTION EXPERIMENT 2

The results of Experiment 1 (as well as Karpicke & Roediger, 2007) indicated that discontinuing test-restudy practice for items that are correctly recalled one time during practice is detrimental for final test performance at a one week retention interval. Conditional analyses provided an initial indication that there is a benefit for correctly recalling items more than one time during practice. However, results also suggested that there is a limit to this benefit.

Similarly, Karpicke and Roediger (2007) conducted conditional post hoc analyses examining final test performance as a function of the number of times items were correctly recalled during practice. The general pattern of their results showed an increase in final test performance as a function of the number of times items were correctly recalled during practice. However, their results also showed a pattern of diminishing returns on final test performance as the number of correct recalls during practice increased.

Although suggestive, an important limitation of the conditional analyses conducted in Experiment 1 and in Karpicke and Roediger (2007) is that items were not assigned to a criterion level of performance during practice (i.e., the number of times an item was correctly recalled during practice was not experimentally manipulated). As a result, differences in item difficulty may have contributed to the results. More specifically, it is likely that items that were correctly recalled multiple times (and
subsequently had a greater likelihood of being correctly recalled at final test) were those items that were normatively easier to recall.

The primary goal of Experiment 2 was to systematically investigate the relationship between the number of correct recalls during practice and subsequent final test performance, removing concerns about item difficulty effects by assigning items to criterion level during practice. Importantly, I also examined this relationship as a function of retention interval. To revisit, the results of Experiment 1 suggest that the apparent inconsistency in results between Pyc and Rawson (in press) and Karpicke and Roediger (2007) was due to differences in the retention interval used in the two studies. To further explore this variable, Experiment 2 included both a short and a long retention interval akin to those used in these two previous studies. Of secondary interest, I also examined the extent to which the relationship between the number of correct recalls during practice and subsequent final test performance depends on ISI during practice.
METHOD

Participants and design. One hundred twenty-nine participants enrolled in Introductory Psychology at Kent State University participated in return for course credit. Criterion level (1, 3, 5, 6, 7, 8, 10) was a within participant variable. ISI (6 versus 34) and retention interval (short RI versus long RI) were between participant variables.

Materials. Materials included 70 Swahili-English translation word pairs (previously normed by Nelson & Dunlosky, 1994). Seven items were assigned to each of ten lists, with an equivalent range of item difficulty within each list. Within each list, one item was randomly assigned to each of the seven criterion levels (randomized anew for each participant); thus, each criterion level had ten items, one from each list. Random assignment of item to criterion level thus minimized concerns about the contribution of item difficulty to effects of criterion level.

Procedure. All task instructions and items were presented via computer. Items received both an initial study trial and test-restudy practice trials. On initial study trials, items received a ten second presentation of both the cue and target (as in Experiment 1). On practice trials, participants were presented with a cue and had eight seconds to recall the target. If the participant recalled the item before eight seconds had elapsed, they could press a “done” key to advance. If an item was correctly recalled, there was no restudy opportunity for the item, based on previous research indicating that feedback does not further improve memory for items that are correctly recalled during practice.
(MacLeod, 1985; Pashler, Cepeda, Wixted, & Rohrer, 2005). Only items that were incorrectly recalled received a four second restudy opportunity before continuing on to the next to-be-learned item.\(^1\) The computer recorded the number of times each item was correctly recalled. Items continued to be practiced until they reached their assigned criterion level of performance (1, 3, 5, 6, 7, 8, or 10 correct recalls). Participants were aware at the outset that they would be tested on items until they reached an “acceptable level of performance”, but were not specifically aware of the number of times items needed to be correctly recalled.

In the ISI 6 group, each of the seven items from the first list were presented for an initial study trial. These items were then presented for test-restudy practice trials until each item reached its assigned criterion level of performance. At that point, the items from the second list were presented for initial study, and so on until all items in each of the ten lists reached their criterion level of performance. Order of list presentation was counterbalanced across participants.

In the ISI 34 group, five of the lists were assigned to the first block of study, and the other five lists were assigned to the second block (with assignment of list to block counterbalanced across participants). All 35 items in the first block were presented for an initial study trial. These items were then presented for test-restudy trials until each item within the first block reached its assigned criterion level of performance. At that point, items from the second block were presented for initial study, and then received test-

---

\(^1\) I decided to make these methodological changes from the last experiment in order to keep the experiment at a reasonable length. More specifically, given the number of items used and the number of times items had to be correctly recalled during practice, I decided to save time where I was able to during the practice phase.
restudy practice trials until each item had reached its assigned criterion level of performance. In both ISI groups, participants were given up to 90 minutes to learn all items to criterion (to keep the length of the experiment manageable).

Following the practice phase, all groups completed a 25 minute reading comprehension filler task that was not related to the main experimental task. Upon completion of the filler task, participants in the short RI group completed the self-paced final cued-recall test for all 70 items. Participants in the long RI group completed the final cued-recall test one week later.
RESULTS AND DISCUSSION

To revisit, the highest level goal of the present experiment was to investigate the relationship between the number of correct recalls during practice and subsequent memory performance. More specifically, I investigated whether there are diminishing returns of retrieval practice, in which correctly recalling an item more times during practice provides no further benefit to final test performance. Percent correctly recalled at final test as a function of group and criterion level is presented in Figure 1.

![Figure 1](image)

Figure 1. Final test performance as a function of the number of times an item was correctly recalled during practice, interstimulus interval (ISI), and retention interval (RI). All standard errors of the mean were between 1.1 and 7.4.
On one hand, the results show a labor and gain effect, providing converging evidence with Experiment 1 and Karpicke and Roediger (2007). One correct recall during practice led to lower levels of final test performance than correctly recalling items more than once, all $t > 2.60$, $p < .01$. The only exception was in the ISI 6 long RI group, in which performance was on the floor; final test performance did not significantly differ after correctly recalling an item once versus ten times during practice. This finding is somewhat surprising and may warrant further investigation in future research; for present purpose, because of the floor effect in this group, further analyses will focus on the other three groups.

On the other hand, the results also show a labor in vain effect. All groups reached an asymptote, in which correctly recalling items more times during practice had no further benefit for final test performance. Additionally, there appear to be differences between the groups in the asymptotic level of performance achieved, as well as the number of correct recalls needed to reach asymptote. To test the extent to which these differences depended on retention interval and ISI, I conducted a series of curve fit analyses. Specifically, I fit the following exponential equation to the data (Busey & Loftus, 1994):

$$p = a \left(1 - e^{-\left(c / r\right)}\right)$$

where $p$ represents the predicted value; $a$ represents the asymptote; $c$ represents criterion; and $r$ represents the rate of approach to asymptote, respectively.2

---

2 This equation also has a $t$ parameter ($c - t$) related to the amount of knowledge people have prior to practice. Due to the nature of my materials it is extremely unlikely that participants have any prior knowledge about the items they are studying. Therefore, I set the $t$ parameter to 0.
First, to examine the extent to which the rate of approach to asymptote differed as a function of retention interval, I compared model fits for the two ISI 34 groups. Specifically, first I fit a *full model* to the data in each group, in which the two parameters \((a\text{ and } r)\) were free to vary for each group. Then I fit a *reduced model*, in which \(a\) was free to vary but \(r\) was constrained to be the same for the two groups. Figure 2 presents the full model (top panel) and reduced model (bottom panel) for each retention interval.

To compare the fit of the two models, I computed an F value, \(F = \frac{\left((\text{RSS}_r - \text{RSS}_f)/H\right)}{\text{RSS}_f/(n-k)}\), where \(\text{RSS}_r\) = residual sum of squares for the reduced model, \(\text{RSS}_f\) = residual sum of squares for the full model, \(n\) = number of observations (14, based on seven means in each group), \(k\) = number of parameters in the full model, and \(H\) = difference between the number of parameters in the full and reduced models. Results indicated that the reduced model yielded a significantly poorer fit to the data, \(F(1,10) = 38.66, p < .001\). Thus, rate of approach to asymptote differed between the short and long RI groups \((r = .724\text{ versus } r = 3.18)\). More correct recalls during practice were needed to reach asymptotic final test performance with a longer versus shorter retention interval.
To examine the extent to which rate of approach to asymptote differed as a function of ISI, I compared model fits for the ISI 6 short RI and ISI 34 short RI groups. Again, in the full model, the two parameters ($a$ and $r$) were free to vary for each group, and in the reduced model, $a$ was free to vary but $r$ was constrained to be the same for the two groups. Results indicated that there was not a significant difference between the full model and reduced model, $F(1,10) = 1.62, p = .232$. Thus, the rate of approach to
asymptote was similar for the two ISI groups. Therefore I only present the full model in Figure 3.

![Figure 3](image)

Figure 3. Mean final test performance for the ISI 6 short RI group and the ISI 34 short RI group as a function of the number of times an item was correctly recalled during practice, and fit of the full model to those data.

Although ISI did not influence the rate of approach to asymptote, ISI did influence the asymptotic level of performance achieved. For this analysis, in the reduced model, \( r \) was free to vary but \( a \) was constrained to be the same for the two groups. Results showed that the asymptotic value estimates for final test performance were significantly higher in the ISI 34 short RI group than in the ISI 6 short RI group (\( a = .648 \) versus \( a = .578 \)), \( F(1,10) = 8.42, p < .05 \). Finally, for purposes of completeness, I also examined asymptote as a function of retention interval. Not surprisingly, the asymptotic values at final test were significantly higher in the ISI 34 short RI group than in the ISI 34 long RI group (\( a = .648 \) versus \( a = .394 \)), \( F(1,10) = 6.33, p < .05 \).
Overall, results from Experiment 2 indicate that more correct recalls are needed during practice for longer retention intervals. However, for both shorter and longer retention intervals, results showed diminishing returns in which there was no additional benefit for final test performance for correctly recalling an item more times during practice.
GENERAL DISCUSSION

Results from two experiments indicated that there is a non-linear relationship between the number of correct recalls during practice and final test performance. Correctly recalling an item more than one time during practice enhanced final test performance (a labor and gain effect); however, there was a limit to the benefit of correctly recalling items more times during practice (a labor in vain effect). Furthermore, the results demonstrate that retention interval influences the rate of diminishing returns from increasing the number of correct recalls during practice, but ISI does not.

The present results are related to findings in the overlearning literature (for a recent review see Driskell, Willis, & Copper, 1992). Driskell et al. (1992) concluded that there is a moderate effect of overlearning; final test performance is higher when items are overlearned during practice compared to when items are not overlearned during practice. However, it is important to note that overlearning has been operationalized in several different ways. Duration based procedures (i.e., Rohrer, Taylor, Pashler, Wixted, & Cepeda, 2005) predetermine the number of practice trials each group receives. For example, the overlearning group may receive ten practice trials, whereas the control group only receives five practice trials. In contrast, in Experiment 2, I manipulated the number of times items had to be correctly recalled during practice rather than using a predetermined number of practice trials (although some previous overlearning studies using duration-based procedures report post-hoc measures to examine the relationship
between final test performance and the number of times items were correctly recalled during practice).

In criterion-based procedures, on the other hand, items are practiced until they reach a specific criterion, after which the amount of additional practice is manipulated. Criterion-based procedures can be broken down further into those in which items receive either fewer or more additional practice trials (e.g., learn items to a criterion of two correct recalls and then receive either ten or 20 additional practice trials, as in Kratochwill, Demuth, Conzemius, 1977), or procedures in which items continued to be practiced until they reach a specified criterion (e.g., 1, 2, 4 correct recalls, as in Nelson, Leonesio, Shimamura, Landwehr, & Narens, 1982). The specific procedure implemented in Experiment 2 is most similar to the latter type of criterion-based procedure. However, only one study (Nelson et al., 1982) has used this particular criterion-based procedure, with the majority of overlearning studies using a duration-based procedure.

Nonetheless, results from this criterion-based overlearning study (Nelson et al., 1982) are consistent with the current results; correctly recalling items more times during practice enhances final test performance compared to correctly recalling items fewer times during practice (a labor and gain effect). However, the current research extends beyond this and other previous overlearning studies in several key ways. First, I included a wider range of criterion levels. For example, Nelson et al. (1982) found a labor and gain effect (as I did for lower criterion levels), but because I utilized a wider range of criterion levels (up to ten), I found a labor in vain effect once higher criterion levels were reached. Second, in the current research I directly manipulated the ISI between practice
trials with items (which to my knowledge has not previously been explored in relation to overlearning). Third, I examined the two aforementioned variables as a function of retention interval. Although previous research has utilized short versus long retention intervals to determine if the benefits of overlearning are maintained over a delay, the relationships between criterion level, ISI, and retention interval have not been adequately explored.

In addition to providing important empirical extensions beyond previous research, the current results also have implications for theories of testing effects. For example, the retrieval hypothesis (discussed in Glover, 1989) states that the extent to which practice tests benefit final test performance depends on the number of complete retrievals during practice. Presumably, items that are retrieved out of short term memory during practice have a lower likelihood of being correctly recalled at final test than items that are recalled from long term memory. In relation to spacing effects, Glover (1989) concluded that differences in final test performance for massed versus spaced presentation schedules are due to the nature of recalls in a massed schedule versus a spaced schedule. Items that are recalled in a massed schedule are likely not complete retrieval events (because they are being recalled from short term memory).

The current results partially support the retrieval hypothesis; in both Experiments 1 and 2, correctly recalling an item more than once during practice enhanced final test performance. However, results from the ISI 6 long RI group in Experiment 2 may provide evidence against the retrieval hypothesis. To revisit, I found no difference in final test performance for items that were correctly recalled one versus ten times during
practice. Although this pattern provides evidence against a retrieval hypothesis, an explanation in line with the retrieval hypothesis could be that retrieval of items during practice in this group were not complete retrieval events. Given that only five items intervened between each practice trial with an item, it is possible that the item was still in short term memory and thus easily recalled during practice. Therefore, practice for items in the ISI 6 group may functionally have been massed.

However, the retrieval hypothesis does not provide a straightforward explanation for the diminishing returns of increasing correct recalls during practice. Better equipped to explain this finding may be another account that has been proposed to explain testing effects, the desirable difficulties account (e.g., Bjork, 1994). According to this account, the extent to which practice tests benefit final test performance depends on the effort expended during practice. When retrieval during practice is effortful, items are more likely to be correctly recalled at final test than items that are effortlessly recalled during practice. In line with this account, one possible explanation for the pattern of results in Experiment 2 (in which final test performance approached asymptote as criterion level during practice increased) is that after a certain number of correct recalls with an item, retrieval was no longer effortful.

Like Verkoeijen, Rikers, and Schmidt (2004) found for spacing effects, there is likely not a single explanation for testing effects, but an interaction between different mechanisms for the testing effects phenomenon. Indeed, the retrieval hypothesis and desirable difficulty account are not mutually exclusive and may both play a role. Importantly, the novel findings of the current research provide constraints for theories of
testing effects, in that any viable theory will need to explain both the labor and gain and the labor in vain effect found here.
FUTURE DIRECTIONS

I am currently planning several follow-up studies to extend the results of the present experiments and to provide a better understanding of the mechanisms underlying testing effects in general and asymptotic levels of performance as a function of criterion level and retention interval in particular. One important account to investigate further is desirable difficulties. More specifically, future studies will assess the retrieval effort expended when items are correctly recalled during practice. To do so, I will record first key press retrieval latencies (i.e., the time it takes participants to begin typing their answer to a target word on any given practice test trial). If retrieval latencies become shorter as the number of correct recalls during practice increases, it is likely that retrieval effort diminishes, which may explain differences in when items reach asymptote.

Although it is expected that all items will eventually reach asymptote, perhaps by counteracting the diminishing retrieval effort during practice I will be able to enhance final test performance levels. That is, participants may benefit more from correctly recalling items when retrieval effort is high as compared to when retrieval effort is low. One way to manipulate the difficulty of retrieval during practice trials is to extend the ISI at the point where retrieval is no longer effortful. For example, if retrieval latencies significantly decrease after five correct recalls with an ISI of 34, extending the next retrieval practice trial to an ISI of 45 may make retrieval of items more effortful, while
still allowing items to be correctly recalled during practice. Thus, the likelihood that an item will be correctly recalled on a later retention test will be increased.

An alternative explanation for the labor in vain effect observed in Experiment 2 concerns contracting lags during practice. Contracting lag refers to the decreasing lag (or ISI) between practice trials as the number of to-be-learned items in a list decreases. For example, the ISI 34 groups initially have an ISI of 34 items between each next practice trial with a given item. As practice test trials continue, more items are being correctly recalled to their criterion level of performance, and subsequently dropped from additional retrieval practice. As this occurs, the items that are left to be recalled to criterion (i.e., those that have been assigned to a higher criterion level) have increasingly short ISIs between each next practice test trial.

A contracting lag between practice trials as items drop out will contribute to retrieval difficulty, but could also have effects on encoding variability. For example, as lag decreases during practice, the encoding variability between each next practice trial with an item decreases because there are fewer items between each next practice trial with a given item. Therefore, it is possible that items reached asymptotic levels of performance when they did for higher criterion levels because of a decrease in encoding variability. I am currently devising methods to differentiate between retrieval difficulty and encoding variability explanations for the current pattern of results.

In Experiment 2, I investigated rate of approach to asymptote for all items as a function of the number of times items were correctly recalled during practice. Another future direction is to investigate differences in rate of approach to asymptote for
individual items as a function of the number of times items are correctly recalled during practice. Perhaps not all items need to reach the same number of correct recalls during practice to achieve the same level of performance on the final test. One possibility is that the criterion level for a given item could be tailored based on an individual’s judgment of learning (JOL). More specifically, on the last correct recall for each criterion level I will ask participants to make a JOL about the likelihood of being able to correctly recall each word pair on a later retention test. In doing so, I will be able to investigate the relationship between the judgments students make about their likelihood of correctly recalling items on a future test, and their actual performance at final test as a function of criterion level. In Experiment 2, it appears that asymptote is reached around five correct recalls. However, participants may be highly confident that they will correctly recall items at final test after only two correct recalls for some items, but after eight correct recalls for other items. First, I will need to determine if students’ JOLs during practice are predictive of final test performance. If so, I may be able cater the number of correct recalls each item receives as a function of JOL magnitude in later studies.

The outcomes of this follow-up research could have important applied implications. Take a student studying for an exam as an example, if students are self testing (i.e., using a flashcard method), when they are highly confident in a correct recall, perhaps they should not continue testing items. In assessing the accuracy of JOLs for predicting final test performance, I may be able to offer students a more efficient way to study for exams. For example, perhaps all items need to be correctly recalled at least two times during practice, but after that point, only items below a certain JOL warrant further
testing. If this is the case, I could limit the amount of time allocated to recalling items judged as well-known, which would leave more time for items students have difficulty learning (or feel they have not learned well).

Of course, it also will be important to generalize results to materials more complex than paired associates. Recent research has suggested that methods for testing items with paired associates do not always generalize to more complex materials, such as key term definition pairs (Rawson, Pyc, Burke, & Dunlosky, in preparation). Using key term definition pairs, Rawson et al. (in preparation) found no difference in final test performance when participants received multiple test-restudy trials versus one test-restudy practice trial. In their second experiment, they found a benefit of multiple test-restudy practice trials only when students engaged in elaborative processing during practice (i.e., comparing and contrasting different key term definition pairs). Future research will need to assess whether the same pattern of results found in Experiment 2 of the current paper, in which items reach asymptote level of performance after correctly recalling items a certain number of times during practice, is also true for more complex materials that students often encounter in their classes.

In sum, the current research makes several important contributions to the testing effects literature. Empirically, the labor and gain but then labor in vain effect presented herein is unique. The finding that there is a limit to the benefit of correctly recalling items more times during practice has not been previously explored in the testing effects literature. The current pattern of data also provides an important foundation for guiding
future research towards a better understanding of the novel patterns shown here, and testing effects more generally.
REFERENCES


Chan, J. C., K., McDermott, K. B., & Roediger, H. L. III (2006). Retrieval-induced facilitation: Initially nontested material can benefit from prior testing of related
material. *Journal of Experimental Psychology: General, 135*, 553-571.


