

ABSTRACT

LAWSON, PATRICK A. Difficulty, Spacing, and the Optimization of Declarative Learning: An Applied Approach. (Under the direction of Dr. Christopher B. Mayhorn.)

Formally learning declarative information is often an arduous, effortful task, as students and language-learners of all ages can attest. While there is an abundance of research into retrieval practice, learning schedules, the spacing effect, and other elements central to the efficient learning of declarative information, these findings can be conflicting or difficult to implement in real-world settings. This is likely due in part to 1) the tendency to study the contributing effects in relative isolation, and 2) placing insufficient emphasis on ecological validity. The experiments described herein aim to generate a more holistic, comprehensive understanding of the factors at play throughout the learning process, their effect sizes, and their relationship to one another. This is accomplished through the comparison of a greater number and variety of learning schedules than prior studies, allowing a broader picture of the effects to emerge.

This work places a particular emphasis on the role of difficulty in declarative learning, revisiting the claims of the desirable difficulty hypothesis. To this end, a novel Composite Accuracy and Subjective Difficulty (CASD) scale is introduced and validated to combine subjective difficulty ratings with objective accuracy to approximate difficulty as a continuous function. This CASD scale outperformed response latency, the most common continuous, retrieval-level proxy for difficulty, by a wide margin. Employing this difficulty scale in conjunction with accuracy measures, multi-level model analyses revealed a sizeable body of evidence against the existence of a desirable difficulty effect. We posit that ease –not difficulty– is beneficial, and that the maximization of ease is central to effective learning. The possibility

that the desirable difficulty effect may be illusory and better attributed to the spacing effect is discussed.

Finally, in a push to apply the previously described findings, a prospective, adaptive algorithm to optimize declarative learning is developed and tested. Most research to date has retrospectively analyzed the efficacy of various types of learning schedules. In a departure from this post-hoc approach, the devised algorithm serves as a proof-of-concept for how one might optimize learning schedules according to the difficulty and accuracy of each retrieval attempt, in real time. This opens the door for learning schedules to be tailored to the precise goals of the learner, while inherently accounting for stimulus difficulty as well as individual differences between learners.

© Copyright 2021 by Patrick Lawson

All Rights Reserved

Difficulty, Spacing, and the Optimization of Declarative Learning:
An Applied Approach

by
Patrick Lawson

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Psychology

Raleigh, North Carolina
2021

APPROVED BY:

Dr. Christopher B. Mayhorn
Committee Chair

Dr. Anne C. McLaughlin

Dr. Jing Feng

Dr. Douglas Gillan

DEDICATION

This work is dedicated to my parents, John and Francesca, my brother, Kevin, and my partner, Alana; your positive influences on me cannot be overstated.

BIOGRAPHY

Patrick Lawson is a doctoral candidate at North Carolina State University in the Human Factors & Applied Cognition program of the Psychology department. Prior to his graduate career at NCSU he received his undergraduate degree from Johns Hopkins University, having majored in Psychological & Brain Sciences. At NCSU his research began with a focus on email phishing susceptibility models, and more recently has centered on the optimization of declarative learning, especially as it relates to the desirable difficulty framework. Outside of his academic studies he has applied his Human Factors background to industry User Experience Research to improve people's interactions with complex systems, primarily in the technology domain.

ACKNOWLEDGEMENTS

This work was made possible in large part by my advisor, Dr. Christopher B. Mayhorn, and I would like to express my deepest gratitude to him for helping me progress as researcher over these past five years. Likewise, my appreciation goes out to my closest friends, whose support was so critical to me. I couldn't have asked for a better group to share the experience with.

TABLE OF CONTENTS

LIST OF TABLES	vii
LIST OF FIGURES	viii
1 Introduction	1
1.1 Declarative Learning	1
1.1.1 Retrieval Practice	1
1.1.2 Learning Schedules	3
1.1.3 Spacing: Total & Relative	4
1.2 Desirable Difficulty	6
1.2.1 Shortcomings of Evidence in Favor	9
1.2.2 Forgetting	10
1.2.3 Availability	11
1.2.4 Tailoring Availability to Learning Goals	13
1.3 Measuring Difficulty	15
1.3.1 Group Level Measurement	15
1.3.2 Retrieval Level Measurement	16
1.4 Gaps in the Literature	17
1.4.1 Schedule Comparisons	17
1.4.2 Performance & Measurement	17
2 Present Study	19
2.1 Overview	19
2.2 Composite Accuracy and Subjective Difficulty (CASD) Scale	21
2.3 Experiment 1	23
2.3.1 Methods	24
2.3.1.1 Participants	24
2.3.1.2 Materials	24
2.3.1.3 Design	24
2.3.1.4 Procedure	26
2.3.2 Results	28
2.3.2.1 CASD Scale Efficacy	28
2.3.2.2 Predicting Ease	30
2.3.2.3 Linear vs. Quadratic Ease Model	31
2.3.2.4 Lookback Window	31
2.3.2.5 All Major Predictors of Ease	34
2.3.2.6 Mediation: Retention Intervals Driving Ease	36
2.3.2.7 Predicting Delayed Retrieval Accuracy	39
2.3.2.8 Relative Spacing: Constant, Contracting, Expanding	40
2.3.2.9 All Major Predictors of Long-Term Retention	43
2.3.3 Discussion: Experiment 1	45
2.3.3.1 CASD Scale Efficacy	46

2.3.3.2	Predicting Ease	47
2.3.3.4	Linear vs. Quadratic Ease Model	48
2.3.3.5	Lookback Window	48
2.3.3.6	All Major Predictors of Ease	49
2.3.3.7	Mediation: Retention Intervals Driving Ease	51
2.3.3.8	Predicting Delayed Retrieval Accuracy	52
2.3.3.9	Relative Spacing: Expanding, Contracting, Constant	53
2.3.3.10	All Major Predictors of Long-Term Retention	54
2.4	Experiment 2	56
2.4.1	Methods	57
2.4.1.1	Participants	57
2.4.1.2	Materials	57
2.4.1.3	Design	57
2.4.1.4	Procedure	59
2.4.2	Results	59
2.4.2.1	Efficacy of Ease Targeting in Adaptive Schedules	60
2.4.2.2	Comparison of Fixed and Adaptive Schedules	62
2.4.3	Discussion: Experiment 2	63
3.	General Discussion	64
3.1	Limitations	70
3.2	Future Directions	72

LIST OF TABLES

Table 1: Lookback Window	33
Table 2: All Major Predictors of Ease	36
Table 3: All Major Predictors of Long-Term Retention	45
Table 4: Schedule Construction: Experiment 2	58

LIST OF FIGURES

Figure 1: Learning Schedule Components	3
Figure 2: Basic Schedule Types	4
Figure 3: Desirable Difficulty Hypothesis Visualization	8
Figure 4: Forgetting Curve, Short-Term	11
Figure 5: Forgetting Curve, Long-Term	11
Figure 6: Availability Curve	12
Figure 7: Doubling Expansion Rate (Expanding)	14
Figure 8: Halving Expansion Rate (Contracting)	14
Figure 9: Composite Accuracy and Subjective Difficulty (CASD) Scale	23
Figure 10: Experiment 1 Stimuli: Prompt and Response	26
Figure 11: Histogram of CASD Responses	28
Figure 12: Schedule Average Accuracy vs. Average Response Latency	29
Figure 13: Schedule Average Accuracy vs. Average Ease (CASD)	29
Figure 14: Lookback Window Effect Visualization	33
Figure 15: Retention Interval Mediation: RI.2Ago Acting Through Ease.1Ago	37
Figure 16: Retention Interval Mediation: RI.3Ago Acting Through Ease.2Ago	38
Figure 17: Retention Interval Mediation: RI.4Ago Acting Through Ease.3Ago	38
Figure 18: Session 1 and Session 2 Performance, by Expansion Type	40
Figure 19: Session 1 and Session 2 Performance, by Schedule	41
Figure 20: Session 1 Accuracy Over Time	42
Figure 21: Session 1 Accuracy of Fixed vs. Adaptive Schedules, per Participant	60
Figure 22: Schedule Accuracy Over Time, Colored by Schedule Type	61
Figure 23: Boxplot of Schedule Performance	62

INTRODUCTION

Declarative Learning

Intentionally learning a new piece of declarative information, such as a fact for an exam or a word in a new language, is often challenging, requiring substantial time and effort to accomplish to a satisfactory degree. There are multitudes of strategies that attempt to make this process easier on the learner, though not all strategies are equally effective. This implies that choosing anything besides the most efficient learning strategy results in wasted time, effort, or both. How, then, can one ensure that they are learning information as efficiently as possible? An abundance of research in this field has identified some methods to get closer to this aim, although large gaps abound. In particular, findings are often difficult to translate to real-world learning applications outside of a laboratory environment. The present work aims to revisit many of these findings with the intent of 1) forming a more holistic understanding of the relevant effects, including how they compare to and interact with one another, and 2) how to leverage them in an applied setting to learn more efficiently. To build a shared framework, let us first review the primary effects known or posited to affect declarative learning.

Retrieval Practice

When studying (i.e., intentionally attempting to learn) a new piece of information it is common to test oneself to measure progress, for instance taking a practice exam to estimate the success of one's studying. Testing in this context is frequently assumed to simply *reveal* the efficacy of prior learning, analogous to the way a computer accesses files without modifying them. In actuality, testing actively *produces* learning gains, rather than only revealing prior learning (Carpenter, 2009). The finding that repeated testing produces greater learning gains than the equivalent amount of time reviewing and restudying that information is referred to as the

testing effect, and has been documented for upwards of a century (Carpenter & DeLosh, 2005; Gates, 1922; Roediger & Karpicke, 2006; Spitzer, 1939). Testing oneself as a means of learning is referred to as retrieval practice (or cued recall) and is generally structured like flash-card study: the learner is given a prompt and attempts to freely retrieve the associated response. Notably, retrieval practice outperforms restudying even when elaborative study is used, a technique that aims to promote effective encoding by integrating the new information into existing schemas (Roediger & Butler, 2011).

The unparalleled efficacy of the testing effect points to the importance of retrievals to the learning process; it seems that the act of retrieval is benefitting the memory in a way that restudy does not. The process of reconsolidation, wherein the act of retrieval temporarily renders the memory more fragile and malleable, has been proposed to play a role in strengthening memories over time as well as allowing them to be updated as necessary (Lee, 2009). Because reconsolidation is thought to occur to a greater extent after retrievals than restudy (Lee, 2009), it may play a role in the observed benefits of retrieval practice as compared to other study techniques. Another potential explanation for the benefit of retrieval practice is the elaborative retrieval hypothesis, which suggests that each retrieval activates “information related to the target response, thereby increasing the chances that activation of any of this information will facilitate later retrieval of the target” (Carpenter, 2009, pp. 1563). These potential mechanisms of action for the testing effect are not mutually exclusive. Regardless of the root cause of the effect, retrieval practice repeatedly produces the greatest learning gains for declarative information, with few exceptions. For this reason, this work solely focuses on learning achieved through retrieval practice paradigms.

Learning Schedules

Retrieval practice, in its simplest form, constitutes one initial learning event where the to-be-learned association is first presented, and at least one subsequent retrieval attempt (which is a learning event in and of itself). Each learning event is separated by a retention interval. These intervals typically refer to the number of intervening, unrelated retrieval attempts, though they may also refer to units of time (or both). We will refer to this entire construct, from the initial learning event to the last retrieval attempt, as a learning schedule –schedule for short. The core components of learning schedules are illustrated in Figure 1.

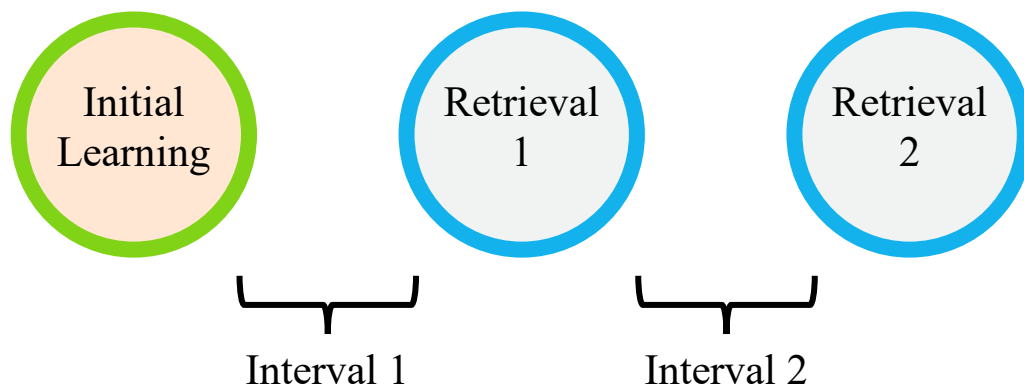


Figure 1. Learning Schedule Components

Learning schedules are defined by 1) the total number of retrieval attempts, 2) the total schedule duration, or time from the initial learning event to the last retrieval attempt, and 3) the relative spacing of the retrieval attempts with respect to one another. In the schedule depicted in Figure 1, for instance, the intervals are identical, making the retrieval attempts equidistant from one another. This is termed a constant schedule. Other schedule types include expanding schedules, where the interval between each retrieval grows over time, or contracting schedules, where the opposite occurs.

The three most commonly referenced schedule types, defined by their relative spacing, are depicted in Figure 2. Although each of these schedule types employs a different relative

spacing pattern, the number of retrieval attempts and the total schedule duration are identical. By definition, the average retention interval is also the same.

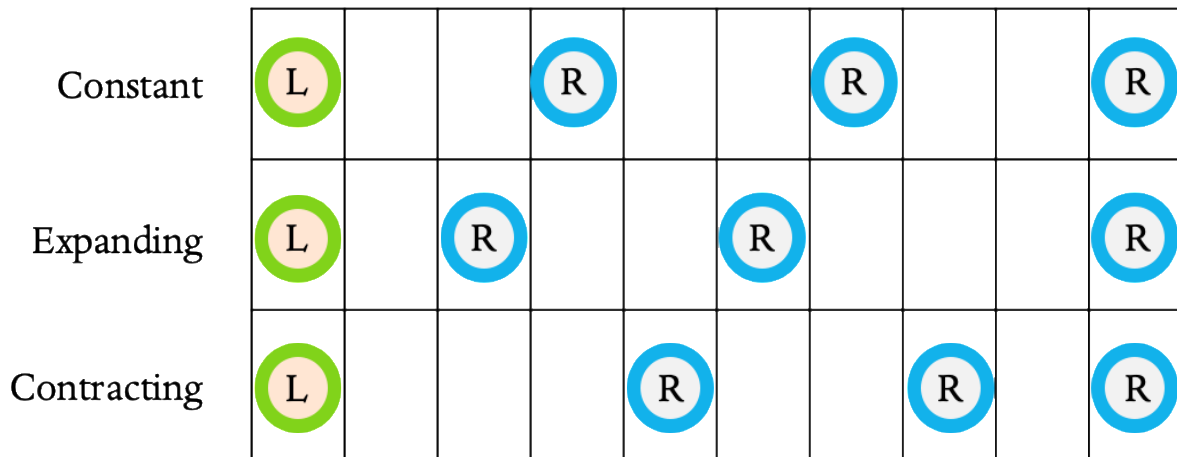


Figure 2. Basic Schedule Types

Spacing: Total & Relative

The spacing characteristics of a learning schedule play a large role in determining the future availability of that memory. When referring to the ‘availability’, or ‘accessibility’ of a memory, we are describing the likelihood of successfully retrieving that memory at a specific point in time (Bjork, 1988; Kang et al., 2014). Low availability equates to low odds of retrieval success. The spacing effect tells us that a schedule’s total duration, or total spacing, positively predicts long-term retention; longer schedules promote longer-lasting learning (Cepeda et al., 2006, Cepeda et al., 2008; Karpicke & Bauernschmidt, 2011).

Somewhat paradoxically, there is often an inverse relationship between the availability of a memory in the short versus the long term. Take, for instance, the different patterns of availability elicited by schedules of short and long total duration (each having the same number of retrieval attempts). Shorter schedules result in increased availability for the earlier retrieval attempts, but decreased availability after a sizeable retention interval. Conversely, longer schedules result in decreased availability for the earlier retrieval attempts, but increased

availability after a sizeable retention interval (Bjork & Bjork, 2011; Bjork, 1988; Landauer & Bjork, 1978). Although it is well established that schedules with long total duration benefit long-term retention, the effect of relative spacing is much less clear. Again, relative spacing refers to the retention interval between retrieval attempts and how that interval changes throughout a schedule (if it changes).

Early research into the effects of relative spacing tended to support the superiority of expanding schedules for long-term retention (Bjork, 1988; Camp, 1996; Landauer & Bjork, 1978). This result was posited to be because the first interval in expanding schedules would be short enough for the information to be highly accessible, making the first retrieval attempt relatively easy. The memory would then be strengthened enough to survive a slightly longer interval, then an even longer interval, and so on, remaining highly accessible throughout (Bjork, 1988). Such an interpretation hinges on the notion that each successive retrieval must be correct to be beneficial (Bjork, 1988). We now know this is not the case, however, especially when feedback is provided. High accuracy at some early stage is not necessarily a prerequisite to achieving high accuracy at some later time (Benjamin et al., 1998). Nonetheless, expanding schedules do have empirical findings backing their superiority over other types of relative spacing, including constant schedules (Kang et al., 2014). Expanding schedules also show unique value in real-world situations, for instance showing strong benefits to dementia patients (Camp, 1996).

Expanding schedules also draw support for their high practical value. Because as the interval between retrieval attempts grows, it is possible to indefinitely add more and more items into the learning rotation. For constant schedules, this would not be possible, as eventually the existing retrieval demands would become overwhelming and prevent more items from being

introduced. Contracting schedules also face this problem, but include the additional issue of having a set “expiration date”. That is, if the retention interval continues to shrink, eventually retrievals will be impractically close to one another, and the schedule will not be able to be continued without modifying the expansion rate. Contracting schedules thus cannot be used in perpetuity.

Regardless of practical consideration, some recent studies have cast doubt on the alleged superiority of expanding schedules. Specifically, there is evidence that expanding schedules are not as effective at promoting long-term retention as equal interval schedules when the retention interval before a delayed test is sufficiently long (Karpicke & Roediger, 2007; Logan & Balota, 2008). While the majority of research agrees that the relative spacing of retrievals plays a role in long-term retention (Karpicke & Roediger, 2007; Logan & Balota, 2008), at least one major study presents evidence that relative spacing has no effect at all. This work instead proposes that only total schedule duration is relevant (Karpicke & Bauernschmidt, 2011).

Desirable Difficulty

In much of the literature concerning retrieval practice, the difficulty of retrievals is often described as a causal factor affecting the accessibility of that information in the future, whether explicitly or implicitly. To exemplify this, let us revisit the comparison between shorter and longer schedules. For longer schedules, it would appear that imposing additional difficulties for the initial retrieval attempts, even to the point of increasing mistakes, leads to improved performance after a sizeable retention interval (Bjork, 1994; Landauer & Bjork, 1978). Put another way, the difficulties imposed at the beginning of the schedule appear to be desirable when the aim is long-term availability. Bjork took this to mean that “*[t]he longer the interval [...], the lower the probability of success, but the greater its benefit for long-term retention*”

(Landauer & Bjork, 1978, pp. 626; Bjork, 1994). The desirable difficulty hypothesis grew from this line of thought, formally stating that 1) difficult and successful retrievals produce greater learning gains than difficult and unsuccessful retrievals, and 2) difficult and successful retrievals produce greater learning gains than easy and successful retrievals (Bjork, 1994; Pyc & Rawson, 2009; Roediger & Karpicke, 2006). The additional difficulties imposed for the early retrieval attempts in the longer schedule, then, were assumed to be directly responsible for the future retrieval success of those items.

The desirable difficulty hypothesis, at its core, proposes that there exists a positive relationship between difficulty and long-term availability, at least for certain levels of difficulty. Following this logic, a schedule that induces some ideal retrieval difficulty at Time(N) may result in an easier retrieval at some future Time(N+X) than a schedule that induces a lower retrieval difficulty at Time(N). This concept is illustrated in Figure 3. In total, the desirable difficulty hypothesis makes three claims, or component hypotheses:

- A) Retrieval difficulty at Time(N) is predictive of retrieval difficulty at Time(N+X).
- B) Increased retrieval difficulty at Time(N) is predictive of decreased retrieval difficulty at Time(N+X).
- C) For some difficulties, a schedule that induces higher difficulty at Time(N) will result in lower difficulty at Time(N+X) than a schedule that induces lower difficulty at Time(N).

Note that it is possible for some of the above component hypotheses to be true while others are false (or that they are all false). Specifically, these can be considered tiered claims, such that one can only be true if the preceding hypotheses are also true. That is, B can only be true if A is true, and C can only be true if both A and B are true. Figure 3. depicts a hypothetical

scenario where all three component hypotheses are true. In the manner that it was originally described, the implication was that all three component hypotheses must be met to constitute a desirable difficulty (Bjork 1994; Bjork 1988; Landauer & Bjork 1978).

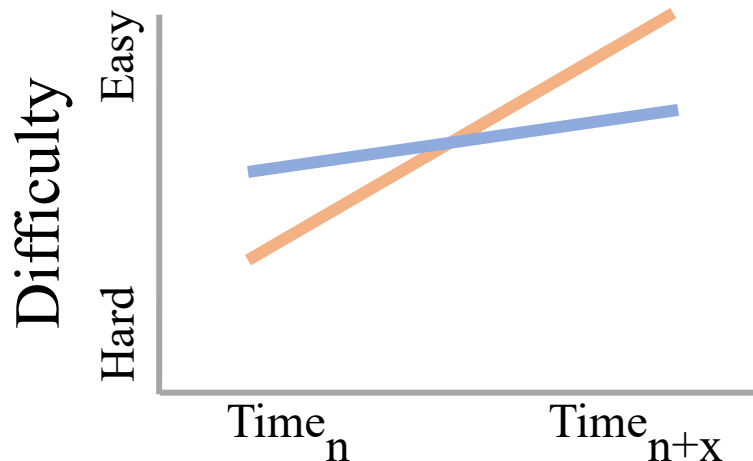


Figure 3. Desirable Difficulty Hypothesis Visualization

There are a few proposed mechanisms of action for the desirable difficulty effect. One proposed mechanism hinges on the notion that high retrieval difficulty may lead to increased processing quality and/or quantity with respect to easier retrievals, which leads to improved retention later on (Bjork, 1994; Gardiner et al., 1973). Another potential mechanism suggests that difficult retrievals may prompt increased encoding effort, and it is this encoding effort that positively affects subsequent retention (Landauer & Bjork, 1978). One final view is that retrieving difficult items is a skill in and of itself. As such, practicing retrieving an item with great difficulty serves as a specific, targeted practice retrieving that item from closer and closer to the brink of being forgotten (Landauer & Bjork, 1978). This view is analogous to practicing escape from hip-deep quicksand: dipping a foot in will not provide adequate practice because it is too easy and therefore dissimilar to the final test.

In this work the focus is primarily on the desirable difficulty effect itself and assessing the conditions (if any) which elicit it, rather than the exact mechanism of action. Nonetheless, the

proposed mechanisms are valuable to consider as they may provide an important lens to interpret findings.

Shortcomings of Evidence in Favor

Though desirable difficulties are often directly mentioned or implied in the literature, relatively few studies claim to have directly tested the hypothesis. Of the few that did, methodological limitations or oddities were common, decreasing the external validity of the findings. For instance, one such study claiming to find support for the desirable difficulty hypotheses elected not to provide feedback, unlike most other studies (Roediger & Karpicke, 2006). Another study used an unrepeated methodology that only concerned itself with correct retrievals rather than total retrieval attempts, and did not control for total schedule duration (Pyc & Rawson, 2009). Yet another found that items retrieved immediately were remembered less well than items that took longer to retrieve, suggesting that the difficulties imposed, as measured by the slower retrievals, enabled them to be better remembered in the future (Gardiner et al., 1973). While this may seem to be strong evidence in favor of the effect, this experiment failed to control for exposure time, such that the items retrieved quickly were available to the subjects for less time than those retrieved slowly, introducing a major confound. Given the varied experimental designs and potential limitations of the previously described studies, it is difficult to place full confidence in their conclusions.

In addition to the relatively small number of studies which claim to directly test the desirable difficulty hypothesis, there are still others which did not set out to specifically test the hypothesis, yet nonetheless invoked desirable difficulty in the interpretation of findings (Carpenter & DeLosh, 2005; Karpicke & Roediger, 2007; Logan & Balota, 2008). Such studies often set out to investigate related phenomena and retroactively concluded (or strongly implied)

that the pattern of their results was likely caused by desirable difficulties. One such study, for instance, set out to investigate the effects of relative spacing on long-term retention. They observed a benefit of equal intervals over expanding intervals for certain learning schedules. This was attributed, post-hoc, to the fact that the first retrieval attempt in the equal-interval schedule occurred later than the first retrieval attempt in the expanding schedule, increasing the difficulty of the first retrieval attempt, thereby causing the item to be better-remembered later on (Logan & Balota, 2008). This study, like many similar ones, was not built to investigate desirable difficulties, yet the hypothesis was invoked to explain the pattern of results. Such interpretations, made after the fact, are scientifically problematic because the experiments were not designed to rule out alternative, competing hypotheses.

Forgetting

We know that the longer an item is not practiced, the more difficult its subsequent retrieval becomes. No discussion of difficulty, then, can be complete without acknowledging the role of forgetting. The fact that unpracticed items become less accessible over time is known as the forgetting curve and has been apparent since Ebbinghaus' foundational research in the late 19th century (Ebbinghaus, 1885). The shape of the forgetting curve has since been determined to closely follow a power function with an asymptote of zero, regardless of the type of stimuli, the retention interval, or whether the success criteria constituted recognition or retrieval (Wixted & Ebbesen, 1991). Examples of this power function of forgetting may be seen in Figures 4 and 5 (reproduced from Wixted & Ebbesen, 1991). Note the similarity of the forgetting curve regardless of the retention interval, or the duration of the study period.

What these graphs (and many others) show us is that the likelihood of correctly retrieving an item from memory decreases steadily over time in a continuous manner, following a power

function. Forgetting is occurring any time learning is not occurring, with the strength of the memory determining the timespan over which it forgotten.

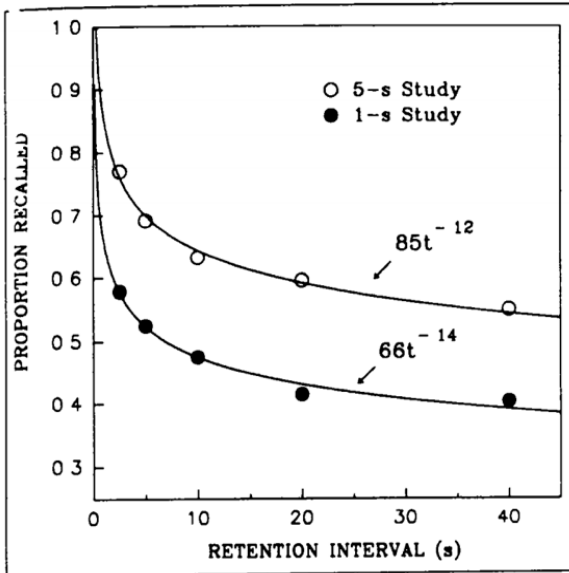


Figure 4. Forgetting Curve, Short Term

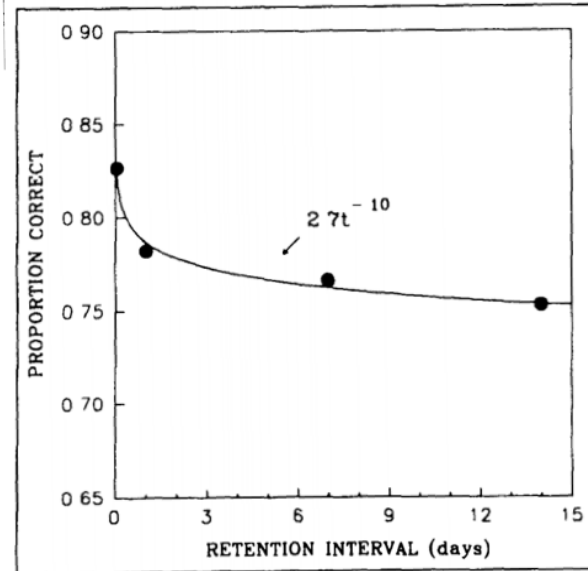


Figure 5. Forgetting Curve, Long Term

Availability

Each time a learning event takes place the memory in question is strengthened and its availability increases. As time passes and forgetting occurs, the availability of the memory decreases, with the speed of the decline determined by the strength of the memory and how well it was previously learned. Heavily practiced items, such as your phone number or postal code, tend to be highly available, easily retrieved, and take a long time to be forgotten. Less practiced items, like the names of new acquaintances, may be less available, more difficult to retrieve, and be forgotten more quickly (Averell & Heathcote, 2011; Wixted & Ebbesen, 1991). Undoubtedly, the availability of a memory changes over time, as it is alternately strengthened through practice and forgotten through disuse.

In Figure 6 the hypothetical availability of an item over time is represented by the purple function. Importantly, assuming feedback is provided, the availability curve returns to maximum

availability after each learning event. This is because providing feedback ensures that inaccurate memories may be updated and corrected, allowing retrieval likelihood to (ideally) reset to a probability of 1 (Lee, 2009; Pashler et al., 2005). As such, we will assume that after each learning event the availability returns to a retrieval likelihood of 100% and declines according to a power function towards an asymptote of 0%.

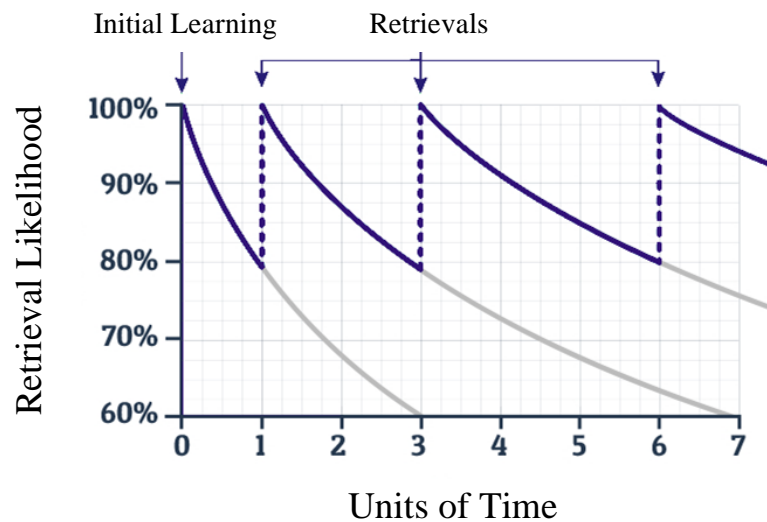


Figure 6. Availability Curve

These opposing forces, learning and forgetting, define the characteristic saw-tooth pattern of the availability curve, spiking upward after learning events and receding through disuse. The gray lines depicted in Figure 6 represent the path the forgetting curves would have taken had the item not been retrieved (or otherwise practiced) again. This is to say that the observed availabilities are a function of when they were retrieved. If any of the retrievals had occurred sooner, they would have occurred closer to the previous retrieval and therefore at a higher point on the availability curve, rendering them more available. Any later, and they would have followed the forgetting curve further down, resulting in a lower availability.

It is also worth noting that after each retrieval attempt the successive forgetting curve becomes slightly less steep, taking greater and greater amounts of time for the memory to decay

to the same availability (80%, in this example). Put another way, each retrieval attempt with feedback produces learning gains, which slow the rate of forgetting (Carpenter, 2009). Per this understanding of the learning and forgetting processes, learning may thus be viewed as the act of making the forgetting curve less steep, thereby staving off forgetting for increasing periods of time.

Tailoring Availability to Learning Goals

In some situations, it may be possible to know exactly when an item will need to be retrieved from memory, and therefore when it is most important that the memory be most available. This might be the case for a student taking a scheduled exam. The student can then tailor their studying to meet their learning goal: for the studied items to have high availability at exam time. In other cases, it may not be possible to predict when an item will next need to be retrieved from memory. A language-learner, for instance, may not know when they will next need to retrieve a particular word from memory. That tricky word that they struggled so hard to learn might never emerge in conversation, or it may turn out to be unexpectedly important.

Given that the language-learner may have vastly different aims than the student preparing for an exam, their ideal learning schedules may look vastly different as well. The language-learner would likely prefer a learning schedule that maintains a moderately high availability throughout, ensuring that they have a good chance at correctly retrieving the word at any given point in time. The student, conversely, may prefer a learning schedule with a very high availability peak, while de-prioritizing availability before and after the exam. These examples demonstrate how the goals of the learner may translate to different ideal patterns of availability, which may in turn be realized by different learning schedules.

To demonstrate how different learning schedules may elicit different patterns of

availability it is useful to look at a few examples drawn from prior research on retrieval practice and the effects of relative spacing (Lawson & Mayhorn, 2019). Figure 7 depicts the accuracy for each retrieval of an expanding schedule where the interval is doubled after each retrieval attempt. Figure 8 shows a contracting schedule where the interval is halved after each retrieval attempt. Figures 7 and 8 are notable in that they depict perfectly inverse schedules; they are the same total duration, have the same number of retrieval attempts, and have an identical pattern of retention intervals, though in the opposite order. That is to say that they are mirror images of one another such that the first retention interval of one schedule is the same as the last retention interval of the other.

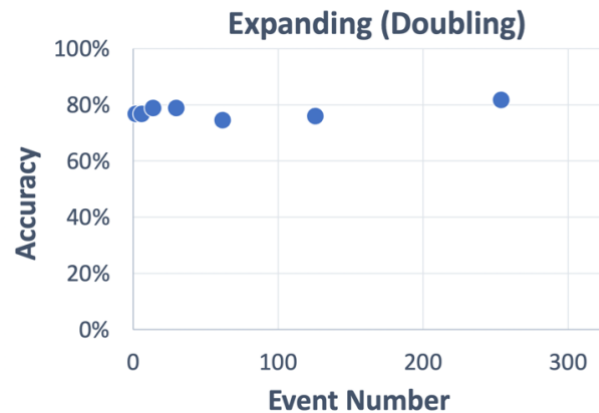


Figure 7: Doubling Expansion Rate (Expanding)

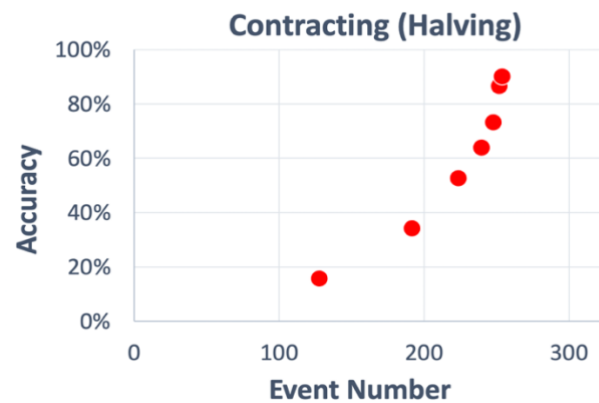


Figure 8: Halving Expansion Rate (Contracting)

As can be seen, the accuracy patterns are vastly different between the expanding and

contracting schedules. Remember that the data points on Figures 7 and 8 only represent the instances where availability (i.e., the likelihood of a successful retrieval) was sampled by means of a retrieval attempt. The full availability curve would look more akin to Figure 6 and demonstrate a saw-tooth pattern, with the plotted points (the retrieval attempts) forming the bottom of each valley, after which the availability rebounds after feedback and begins decaying according to the new forgetting curve. Notably, this expanding schedule produced similar average retrieval accuracies for each retrieval: roughly 80%. In the contracting schedule, the average accuracy of the first retrieval was below 20%, while the last was above 90%. Such a contracting schedule might then be beneficial for the student taking a scheduled test, as it yields the highest availability peak, and does so at the end of the schedule (i.e., at exam time). The language-learner may prefer the more consistent availability pattern of the expanding schedule, even if the maximum realized availability is slightly lower than the corresponding contracting schedule.

Measuring Difficulty

Group Level Measurement

To this point difficulty has not been operationally defined, and indeed is only infrequently defined in the literature. Often it is used interchangeably with accuracy when describing the aggregate performance of a group (Kang et al. 2014; Karpicke & Roediger 2007; Landauer & Bjork 1978; Roediger & Karpicke 2006). In this way aggregate accuracy may be used to describe how a certain learning schedule tends to produce increased or decreased difficulty where accuracy is taken to be inversely proportional to difficulty (that is, high accuracy = low difficulty). Using average accuracy as a measure of difficulty is valuable when describing aggregated trends across many participants, though it is of little value when investigating

difficulty at the level of a single retrieval. Similarly, this type of aggregated accuracy measure can only be analyzed retrospectively, greatly reducing its value in applied, forward-facing use cases.

Retrieval Level Measurement

There are three primary ways to measure or estimate difficulty at the retrieval level:

- 1) Retrieval accuracy
- 2) Response latency
- 3) Subjective rating from participant

Retrieval accuracy is naturally dichotomous, able to be correct or incorrect, yet few would argue that all successful retrievals are equally difficult. Additionally, it is well-documented that forgetting occurs according to a power function, in a continuous manner (Averell & Heathcote, 2011; Wixted & Ebbesen, 1991). As such, valuable data are lost by measuring a continuous process (forgetting) using a binary outcome measure (accuracy).

Response latency is the most common of these three measurement methods (Benjamin et al., 1998; Gardiner et al., 1973; Karpicke & Roediger, 2007; Pyc & Rawson, 2009). It is built on the premise that more difficult retrievals should be less fluid and therefore take longer than easier retrievals (Gardiner et al., 1973; Karpicke & Roediger, 2007; Pyc & Rawson, 2009). While this measure has been validated as being correlated with accuracy, outside of a controlled laboratory setting it may be problematic to assume that all delays in responding are solely attributable to difficulty. To give an example, sneezing before making a response does not necessarily mean the retrieval was difficult.

Subjective measures have also been used to measure difficulty, for instance to have participants rate whether retrievals were easy or difficult, and whether they elicited the tip-of-

the-tongue phenomenon (Gardiner et al., 1973). Such subjective measures are the least commonly used proxy for difficulty in the literature.

Gaps in the Literature

Schedule Comparisons

One shortcoming of previous studies is the relatively low number and type of schedules that tend to be compared to one another. The majority of studies compare a single-digit number of schedules, and often just a handful (Balota et al., 2006; Carpenter & DeLosh, 2005; Cull, 2000; Logan & Balota, 2008). Such approaches restrict the scope of the findings, as the dimensions along which the schedules vary must be correspondingly limited. Given these limitations, such studies cannot simultaneously compare the influences of the variety of factors hypothesized to impact availability, including at a bare minimum: total schedule duration, relative spacing, number of retrieval attempts, and retrieval difficulty. Comparing a greater gamut of schedules will allow for a more holistic, comprehensive analysis of how the various factors contribute to learning (if they indeed contribute), and how they relate to one another.

Performance & Measurement

Another broad shortcoming of previous research is the choice of how to measure 1) the efficacy of a certain type of learning schedule, as well as 2) the difficulty of retrievals within them. Regarding the measurement of schedule efficacy, success is often taken to mean the availability (or retrieval likelihood) on a single delayed test of retention after some pre-specified interval. These studies tend to compare and contrast performance at some early point, often termed the ‘practice’, ‘learning’, or ‘acquisition’ phase, to performance at some later point - the ‘retention test’. Almost all emphasis is placed on availability at the time of the retention test, where earlier availability is largely viewed as simply a means to produce success on this delayed

test. This strict division between a ‘practice’ period and some ‘retention test’ seems unwarranted, given that each of the retrieval attempts, including the final retention test, are structured identically. Instead, it is the authors’ view that each retrieval attempt simply constitutes a piece of the ever-growing learning schedule.

In this view, the performance of a learning schedule should be assessed according to the sum pattern of availability it elicits and its performance across all retrieval attempts. This is not to say that success on each and every retrieval attempt must be equally weighted and prioritized. Rather, it is to highlight that the distinction between retrievals during a “learning period” and those during a final “retention test” is a false dichotomy; they are all learning events that constitute the greater learning schedule. We are aware of only one study that operationalized success as the sum availability throughout a learning schedule’s entire duration, thereby considering all retrieval attempts (Kang et al., 2014).

Having (briefly) considered the criteria on which to measure a schedule’s efficacy, let us turn to the measurement of difficulty. Difficulty is most frequently defined as average accuracy, where binary retrieval success is aggregated across many participants. This allows researchers to say things like “schedule X resulted in a 10% success rate on Y retrieval attempt, evidence of high difficulty”. While useful for describing general difficulty trends, this approach does not allow difficulty to be captured as a continuous function at the level of each individual retrieval; a retrieval may only be correct or incorrect. If difficulty is truly involved in learning, then it should apply to individual retrievals, such that an individual’s prior retrieval difficulties (captured as a continuous variable) should be shown to influence future retrieval difficulties. This cannot be done with aggregated measures of binary measurements. Additionally, aggregated approximations of difficulty must necessarily be analyzed retrospectively, after the fact. This

precludes their use in any sort of flexible, adaptive learning schedule that operates in real time, rather than simply following a predetermined retrieval pattern.

To capture difficulty as a continuous process at the retrieval level many studies have used response latency as a proxy for difficulty (Benjamin et al., 1998; Gardiner et al., 1973; Karpicke & Roediger, 2007; Pyc & Rawson, 2009). As there are a number of factors that may influence response speed, it is warranted to revisit the relationship between latency and difficulty and assess the strength of this correlation. Subjective retrieval difficulty has also been used in the past to capture difficulty at the person level, though exceedingly rarely (Gardiner et al., 1973). Given the naturally subjective component of difficulty, as well as the paucity of this approach in the literature, it is prudent to probe the utility of direct, subjective estimates of difficulty.

The present study will thus sample 1) retrieval latency, 2) subjective retrieval difficulty, and 3) accuracy for each retrieval attempt. Measuring all three of these upon each retrieval attempt will allow for the direct comparison of each approach in estimating retrieval difficulty as well as the probability of success. To our knowledge, such a multi-pronged approach has not yet been employed and may constitute a methodological contribution to the literature.

PRESENT STUDY

Overview

The overarching aims of this work were tripartite. The first was to explore and validate the utility of a Composite Accuracy and Subjective Difficulty (CASD) scale in the continuous approximation of retrieval difficulty, especially as compared to existing methods. The second was to shed light on the factors affecting the efficacy of retrieval-based learning schedules in greater detail than previous works, employing the CASD scale to do so. The third major aim was to leverage these findings in a highly applied setting to adaptively support learning goals – in real

time – and target specific retrieval difficulties.

To accomplish these aims the present work consisted of two experiments. In the first experiment, a multitude of learning schedules were compared to one another to investigate a variety of factors proposed to affect retrieval likelihood in both the short and long term. These factors included total spacing, relative spacing, retrieval attempts, and retrieval difficulty, among others. In particular, the desirable difficulty hypothesis was tested more rigorously than prior research.

To test the desirable difficulty hypothesis, a novel Composite Accuracy and Subjective Difficulty (CASD) scale was utilized, following initial success in previous exploratory research. This scale considers both retrieval accuracy and subjective difficulty to estimate retrieval difficulty at the level of individual retrievals, rather than an aggregated level. Exploratory work showed this CASD scale to explain 50% more variance than response latency (Lawson & Mayhorn, 2019). This is notable given that response latency is the most commonly used proxy for difficulty (Gardiner et al., 1973; Pyc & Rawson, 2009).

To assess the role of difficulty in learning efficacy, expanding, contracting, and constant interval schedules were compared, in greater number and with greater diversity among them than previous works. Notably, each trio of an expanding, contracting, and constant schedule were matched with respect to retrieval attempts as well as total spacing. In addition to providing direct evidence (rather than post-hoc assertions) for or against the desirable difficulty framework, this paradigm has the added benefit of testing the effect of relative spacing, along with many other potentially relevant variables.

In the second experiment, an algorithm was employed to maintain approximately constant retrieval difficulties throughout the learning schedule by adaptively determining the

ideal length of the next retention interval based on prior retrieval difficulties. Regardless of evidence for or against the desirable difficulty framework found in the first phase of the study, the maintenance of a constant retrieval difficulty is nonetheless valuable. Most importantly, an adaptive scheduling algorithm establishes a proof-of-concept for how one might tailor schedules to meet specific learning goals. This type of adaptive schedule also has the added benefit, in theory, of being able to automatically account for the difficulty of the to-be-learned information, as well as the abilities of the learner.

Composite Accuracy and Subjective Difficulty (CASD) Scale

This study expanded on previous uses of subjective retrieval difficulty (Gardiner et al., 1973), with major updates. Both experiments 1 and 2 employed this novel CASD scale and recorded this value for each retrieval attempt, in addition to more established measures such as response latency (Gardiner et al., 1973; Pyc & Rawson, 2009) or accuracy.

In one of the earliest utilizations of the subjective retrieval difficulty, Gardiner et al. captured subjective difficulty estimates *before* retrieval, with participants rating how close or far they felt from retrieving the correct response (Gardiner et al., 1973). The present rating, conversely, is captured *after* the retrieval attempt, allowing for a retrospective judgment of difficulty to be made. This decision was made to not interrupt the retrieval process by requiring the participant to estimate the probability of success before making a response, which would preclude the collection of response latency. By having the participant make a difficulty judgment *after* each retrieval, we were able to capture both subjective difficulty and response latency for each retrieval, in a way that they may be later compared.

Notably, in this paradigm participants are never asked to estimate the probability of the previous retrieval having been successful – they are instead only asked about their subjective

experience of the retrieval. This decision was made with the intention of avoiding the “knew-it-all-along” fallacy, a tendency towards revisionism where knowledge of the outcome biases participants’ estimates of their prior likelihood of success (Wood, 1978). If the retrieval was correct, participants were asked how “difficult or easy” the retrieval felt. If the retrieval was incorrect, participants were asked how “far or close” they felt from generating the correct response.

Two reasons prompted the decision to bifurcate the scale according to the accuracy of the response. The primary reason was to adhere most closely to the claims of the desirable difficulty hypothesis to best test them. The claims of the desirable difficulty hypothesis are that 1) difficult and successful retrievals produce greater learning gains than difficult and unsuccessful retrievals, and 2) difficult and successful retrievals produce greater learning gains than easy and successful retrievals (Bjork, 1994; Pyc & Rawson, 2009; Roediger & Karpicke, 2006). Because the desirable difficulty hypothesis references both the accuracy of the retrieval as well as its subjective difficulty, this CASD scale was determined to be the method best suited to evaluate both claims.

A secondary reason to bifurcate the subjective scale according to accuracy was to reduce the complexity of the question posed to participants. The alternative, asking the participant the same question regardless of response accuracy, was avoided over concern that such a question would likely directly require asking about the probability of the previous retrieval having been successful. Such a question would have to be relatively complex and cumbersome, and responses would presumably be more susceptible to the “knew-it-all-along” fallacy.

For both the correct and incorrect responses, a 7-point Likert scales was used. For both scales, lower numeric responses corresponded to greater subjective difficulty. These scales were

then recoded and combined to form a semi-continuous 14-point scale from unsuccessful retrievals that felt far from being correct, to successful retrievals that felt easy. For more direct comparability with accuracy, the scale was framed around ease rather than difficulty. That is, the highest ease (or lowest difficulty) was placed at the top of the scale in numerical value and position (as illustrated in Figure 9). Time is included on the y-axis to be able to observe changes in difficulty as a longitudinal process (Lawson & Mayhorn, 2019).

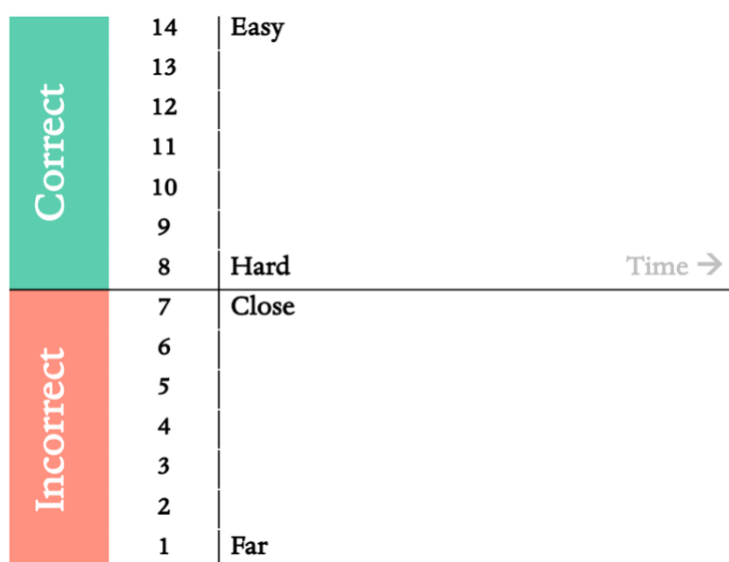


Figure 9. Composite Accuracy and Subjective Difficulty (CASD) Scale

In a prior preliminary study, this subjective retrieval difficulty scale explained 88% of the variance in a schedule's average retrieval accuracy (Lawson & Mayhorn, 2019). This was a full 50% higher than the 38% of variance explained by response latency (Lawson & Mayhorn, 2019). The present study seeks to confirm the strong relationship observed between this CASD value and response accuracy, among other aims.

Experiment 1

In this two-part experiment, participants attempted to learn 27 associated pairs according to 27 unique learning schedules. The associated pairs consisted of a 3x3 black and white

checkered grid (the prompt), and a 3-digit number (the response). These associations were learned via schedules consisting of between 3 and 11 retrieval attempts over the course of nearly 2 hours. This period is designated Session 1. Then, 3 days later, participants performed one final retrieval attempt per schedule, to assess long-term retention. This is designated Session 2.

Method

Participants

Seventy-eight undergraduate students at least 18 years of age enrolled at a large university in North Carolina participated in this 2-part experiment and were compensated with course credit.

Materials

The experimentation platform Psytoolkit was used to build and subsequently host the experiment (Stoet, 2010; Stoet, 2017). After completion of the Session 1, participants were contacted through their university email to 1) inform them of when they were scheduled to complete Session 2, and 2) provide them the URL to access Session 2 at the designated time.

Design

A total of 27 schedules were compared concurrently, rather than sequentially, following previous research (Carpenter & DeLosh, 2005; Karpicke & Roediger, 2007; Karpicke & Bauernschmidt, 2011). These schedules were staggered and interlocked in a way that no two learning events were scheduled to occur at the same time. Nine were expanding schedules, nine were contracting, and nine were constant. Importantly, each schedule was unique, derived from all possible combinations of 3 factors with 3 levels each.

This experiment thus employed a 3x3x3 study design: *3 schedule types* (expanding, contracting, constant) x *3 expansion rates* (Fibonacci, doubling, tripling) x *3 first retention*

intervals. The Fibonacci expansion rate was chosen because it expands more slowly than a doubling schedule while resulting in integer intervals, rather than any specific hypothesis regarding supposed benefits of that exact rate. The first retention intervals and expansion rates were used to define the expanding schedules, which were in turn used to derive the corresponding contracting and constant schedules.

Let us elaborate on this schedule derivation process. First, the expanding schedules were derived according to all combinations of the 3 expansion rates (Fibonacci, doubling, and tripling expansion) and 3 first retention intervals (1, 6, or 11 units of time). This 3x3 cross results in 9 total expanding schedules. The 9 contracting schedules were then defined as the mirror images of the expanding schedules. That is, each of the retention intervals was the same length, but they occurred in reverse order. Finally, the constant schedules were derived by finding the average retention interval of the above expanding and contracting schedules. In cases where this value was not a whole number, retention intervals of the two nearest whole numbers were used in a proportion yielding the correct average interval.

Each of these manipulations was within-subject, such that each participant completed all 27 schedules derived from the 3x3x3 crossing of conditions. This 3x3x3 study design was selected to yield a wide distribution of schedules, systematically varying along three major conditions: schedule type (expanding, constant, contracting), expansion rate, and first retention interval. In this manner we generated a broader array of schedules than previous research in order to form a more holistic understanding of the aspects of learning schedules which most affect the learning process, either positively or negatively.

Procedure

Participants accessed the study via SONA, the university's experimentation management

site, and elected to participate after reading a description of the experiment. This study was administered remotely, entirely online, and participants completed the study on a computer of their choosing (not provided). Participants were not permitted to take major breaks once they had initiated the experiment in their internet browser, though there were extended periods requiring little mental effort. This first portion of this experiment (Session 1) required approximately 2 hours.

Prior to beginning Session 1 of the experiment participants were given a brief tutorial and completed 10 practice trials to familiarize themselves with the task. The experiment was divided into 360 discrete events, falling into 3 categories: an initial learning event, a retrieval attempt, or a working memory task. Each of these events occupied an 18-second block of time. Descriptions of each are provided below.

- **Initial learning event:** The participant was instructed to learn the number associated with a 3x3 grid of randomly generated white or black tiles. This number was 3 digits long and was also randomly generated. An example is shown in Figure 10. After 10 seconds of learning the association a Judgment of Learning (JOL) estimate was made (Dunlosky &

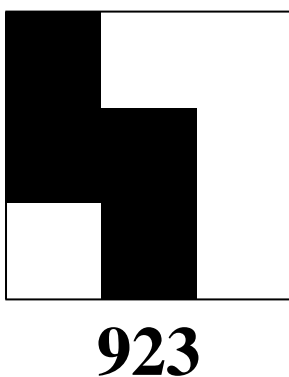


Figure 10. Experiment 1 Stimuli: Prompt and Response

Nelson, 1992). In this JOL, participants were prompted to answer the question “How well do you believe you have learned this association?” responding according to a 7-point

Likert scale, where 1 denoted items learned poorly and 7 denoted items learned well.

- **Retrieval attempt:** The participant was prompted with a 3x3 grid and asked to enter the corresponding numerical response using their keyboard. They had 10 seconds to begin their response. After making (or not making) a response, feedback was provided and the correct response was displayed. If the response was correct, participants responded to a 7-point Likert scale asking, “How easy or hard was it to remember the number?” If the response was incorrect, participants responded to a 7-point Likert scale asking, “How close or far did you feel you were from remembering the number?” Participants had 5 seconds to respond to the Likert scale.
- **Working Memory task:** Any event space not occupied by an initial learning event or a retrieval attempt was occupied with a Working Memory task. Participants were asked to count the number of dots presented on screen, which varied randomly between 6 and 9 (inclusive). The participant then held the answer in memory for 5 seconds, after which they were prompted to enter their response.

Participants attempted to learn all 27 paired associates according to the 27 unique learning schedules, performing Working Memory tasks for all events not occupied by a retrieval attempt or an initial learning event. This was done to prevent the active rehearsal of associations.

Upon completion of Session 1, participants were scheduled to complete Session 2, which occurred 3 days after the completion of Session 1. At the specified time, participants received a URL link to Session 2 at their university email address. This email was delivered 2 days and 18 hours after the completion of Session 1. Participants were then given 12 hours to begin Session 2, which required roughly 15 minutes to complete. Session 2 consisted of a single retrieval attempt for each of the 27 paired associates learned throughout the 27 schedules. Participants

were then thanked for their time and assigned course credit.

Results

CASD Scale Efficacy

As a preliminary overview of responses, we looked at a histogram of all CASD values collected (Figure 11). We can see that responses were not normally distributed and were skewed towards the extremes of the scale. This is not necessarily problematic or unexpected, although it suggests there may be room for refinement in future scales.

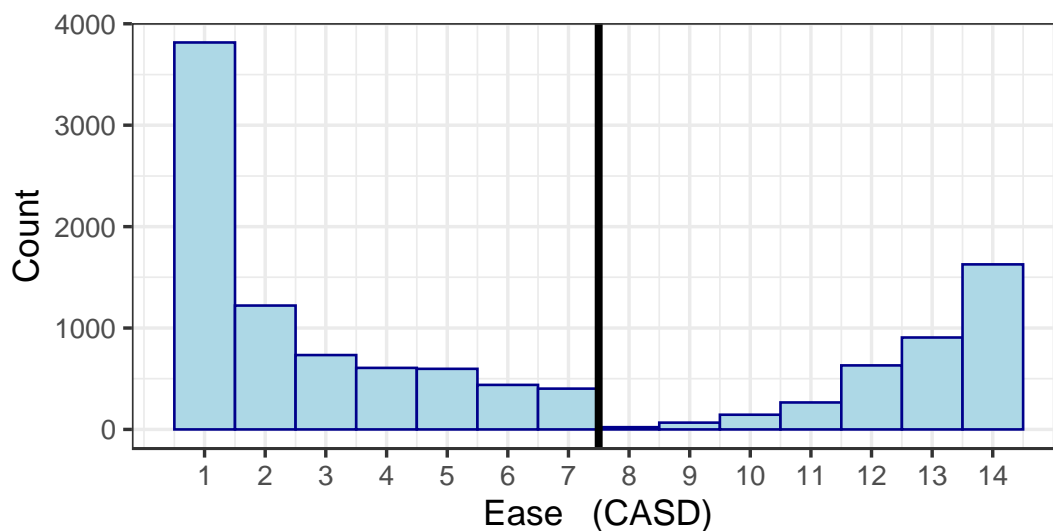


Figure 11: Histogram of CASD Responses

Each participant's retrieval accuracy, response latency, and CASD ease ratings were calculated for each of the 27 schedules. These data were aggregated by participant-schedule, such that each participant generated one data point per schedule (for each of the three factors investigated: retrieval accuracy, reaction time, and CASD ease ratings). Average accuracy was

negatively correlated with response latency ($r = -.36, p < .001$), and positively correlated with

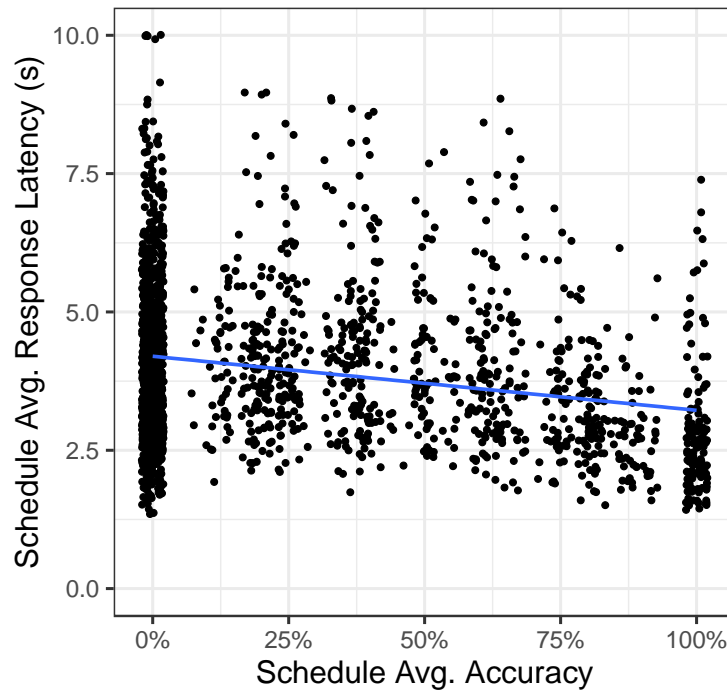


Figure 12: Schedule Average Accuracy vs. Average Response Latency

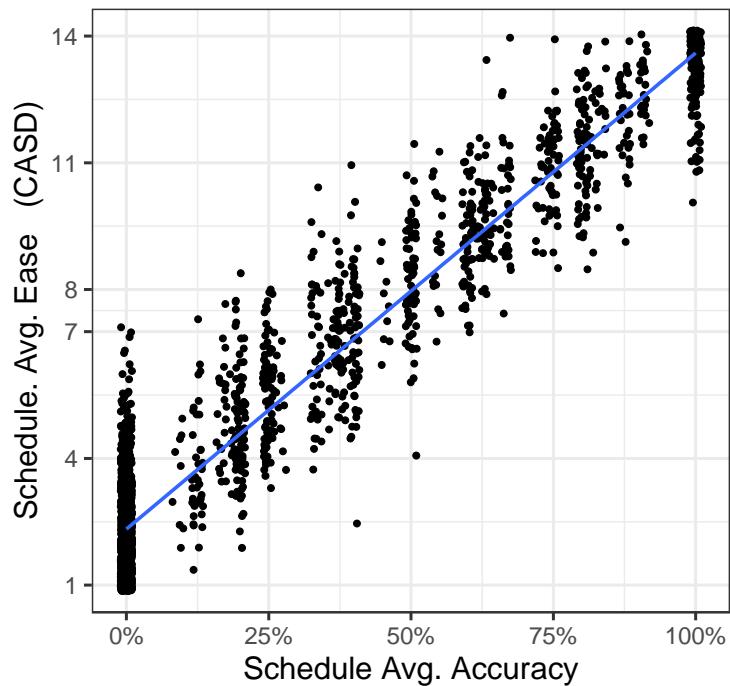


Figure 13: Schedule Average Accuracy vs. Average Ease (CASD)

subjective difficulty rating ($r = .91, p < .001$). The CASD scale explained 83% of the variance in

schedule response accuracy, a full 70% more than response latency, which explained 13% of variance in schedule response accuracy. Figure 12 shows the relationship between schedule accuracy and schedule response latency, and Figure 13 show the relationship between schedule accuracy and schedule ease, as measured by the CASD scale.

Predicting Ease

Before beginning more complex hierarchical analyses a preliminary analysis was conducted to ensure that there was sufficient variability at Levels 1, 2, and 3 to warrant continuation with a multi-level approach (Nezlek, 2001; Raudenbush & Bryk, 2002). This preliminary analysis is termed a fully unconditional model (also referred to as a null model), in which no term other than the intercept is included at any level (Nezlek, 2001). The Ease of each retrieval, as determined by the CASD scale, was used as the Dependent Variable (DV).

- Level 1 (Retrieval)

- $Ease_{rsp} = \beta_{0sp} + e_{rsp}$

- Level 2 (Schedule)

- $\beta_{0sp} = \gamma_{00p} + u_{0sp}$

- Level 3 (Participant)

- $\gamma_{00p} = \delta_{000} + v_{00p}$

Results from this analysis indicated that 71% of the variability in retrieval Ease was at Level 1, the retrieval level, ($\sigma_e^2 = 18.83$), 10% was at Level 2, the schedule level, ($\sigma_u^2 = 2.58$), and 19% was at Level 3 ($\sigma_v^2 = 5.01$), the participant level. Therefore, the fully unconditional model indicated that there was sufficient variability for continued analyses. Further, this 3-level model was found to explain significantly more variance than the equivalent 2-level model, $\chi^2(3,4) = 2413.3, p < .001$.

Linear vs. Quadratic Ease Model

Having established the validity of hierarchical modeling and the 3-level structure of the data, we turned to the question of whether Ease was best modeled as a linear or quadratic term. This was done by comparing two distinct models where the only predictor in each model was Ease. The linear version of these comparison models follows.

- Level 1 (Retrieval)

- $Ease_{rsp} = \beta_{0sp} + \beta_{1sp} (Ease1Ago_{rsp}) + e_{rsp}$

- Level 2 (Schedule)

- $\beta_{0sp} = \gamma_{00p} + u_{0sp}$

- $\beta_{1sp} = \gamma_{10p}$

- Level 3 (Participant)

- $\gamma_{00p} = \delta_{000} + v_{00p}$

- $\gamma_{10p} = \delta_{100}$

The quadratic model was identical, except that the *Ease1Ago* term was raised to the 2nd power ($Ease1Ago_{rsp}^2$). This analysis revealed that more variance was explained modeling Ease linearly than quadratically $\chi^2(5,5) = 199.2, p < .001$.

Lookback Window

Next, we investigated the influence of the ease of previous retrievals on the ease of the current retrieval, here referred to as a lookback window. This was designed to explore a key tenet of the desirable difficulty hypothesis, namely whether increasing the difficulty of a prior retrieval may lead to decreased difficulty for the present retrieval. To do this we fitted a linear mixed model to predict *Ease* with predictors *Ease1ago*, *Ease2ago*, *Ease3ago* and *Ease4ago*. Random effect error terms were included at each level. The full model formula is provided below.

- Level 1 (Retrieval)
 - $Ease_{rsp} = \beta_{0sp} + \beta_{1sp} (Ease.1Ago_{rsp}) + \beta_{2sp} (Ease.2Ago_{rsp}) + \beta_{3sp} (Ease.3Ago_{rsp}) + \beta_{4sp} (Ease.4Ago_{rsp}) + e_{rsp}$
- Level 2 (Schedule)
 - $\beta_{0sp} = \gamma_{00p} + u_{0sp}$
 - $\beta_{1sp} = \gamma_{10p}$
 - $\beta_{2sp} = \gamma_{20p}$
 - $\beta_{3sp} = \gamma_{30p}$
 - $\beta_{4sp} = \gamma_{40p}$
- Level 3 (Participant)
 - $\gamma_{00p} = \delta_{000} + v_{00p}$
 - $\gamma_{10p} = \delta_{100}$
 - $\gamma_{20p} = \delta_{200}$
 - $\gamma_{30p} = \delta_{300}$
 - $\gamma_{40p} = \delta_{400}$

The model's total explanatory power was substantial (conditional $R^2 = 0.63$) and the part related to the fixed effects alone (marginal R^2) was 0.57. The model's intercept was at 1.73 (95% CI [1.24, 2.23], $t(3288) = 6.89$, $p < .001$). The effects of *Ease.1Ago*, *Ease.2Ago*, and *Ease.3Ago* were all significantly positive at the $p=.001$ threshold, while *Ease.4Ago* was significant at the $p=.05$ threshold. Table 1 illustrates these findings in greater depth.

These results demonstrate that the ease of each of the previous four retrieval attempts positively and uniquely predicted current retrieval ease. Figure 14 illustrates the size and directionality of these effects.

Table 1. Lookback Window

Ease			
<i>Predictors</i>	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	1.73	1.24 – 2.23	< 0.001
Ease.1Ago	0.50	0.47 – 0.54	< 0.001
Ease.2Ago	0.21	0.17 – 0.25	< 0.001
Ease.3Ago	0.07	0.03 – 0.11	< 0.001
Ease.4Ago	0.04	0.00 – 0.07	0.043

Random Effects	
σ^2	11.12
$\tau_{00 \text{ par}}$	0.66
$\tau_{00 \text{ sch}}$	0.85
ICC	0.12
N_{sch}	21
N_{par}	78
Observations	3296
Marginal R^2 / Conditional R^2	0.575 / 0.626

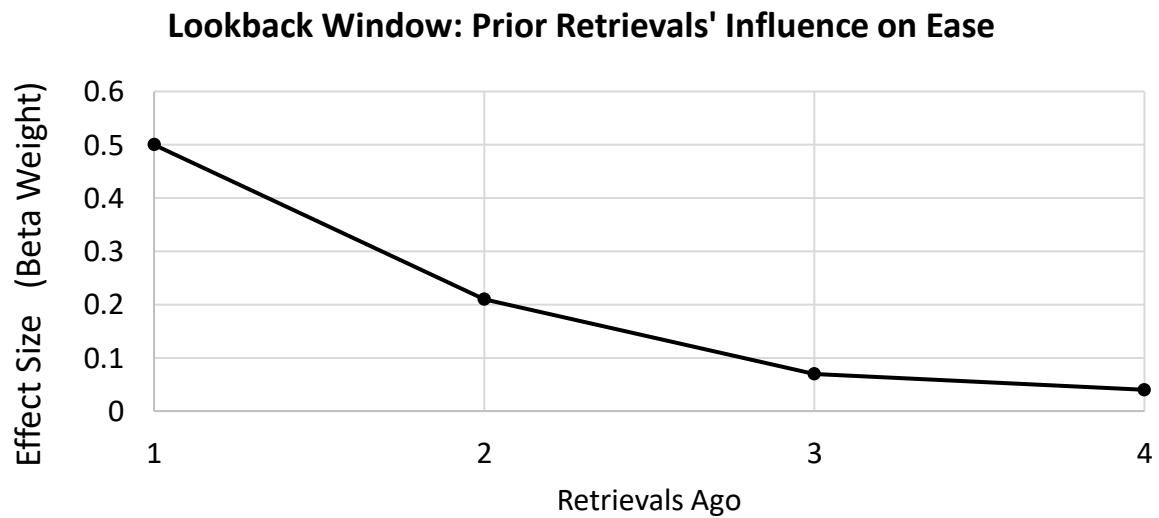


Figure 14. Lookback Window Effect Visualization

All Major Predictors of Ease

Following evidence for the strong, positive predictive power of the ease of previous retrievals, we sought to understand how to best predict retrieval ease. This involved creating a large model including all major predictors thought to influence retrieval ease. These candidate predictors were selected to concurrently assess a multitude of factors thought to affect retrieval ease, drawn both from the broader literature and this work's specific hypotheses. To assess these factors we fitted a linear mixed model to predict *Ease*, and included the following predictors in the model: *Ease.1Ago*, *Ease.2Ago*, *Ease.3ago*, *RI.1Ago*, *Retrieval.Number*, *Duration.per.Retrieval*, *Schedule.Expansion.Rate*, *1st.Retention.Interval*, *Judgment.of.Learning* and *P.Average.Accuracy.Session1*. Random effect error terms were included at each level. The full model formula is provided below.

■ Level 1 (Retrieval)

- $$Ease_{rsp} = \beta_{0sp} + \beta_{1sp}(Ease.1Ago) + \beta_{2sp}(Ease.2Ago) + \beta_{3sp}(Ease.3Ago) + \beta_{4sp}(RI.1Ago) + \beta_{5sp}(Retrieval.Number) + e_{rsp}$$

■ Level 2 (Schedule)

- $$\beta_{0sp} = \gamma_{00p} + \gamma_{01p}(Duration.per.Retrieval) + \gamma_{02p}(Schedule.Expansion.Rate) + \gamma_{03p}(1^{st}.Retention.Interval) + \gamma_{04p}(Judgment.of.Learning) + u_{0sp}$$
- $$\beta_{1sp} = \gamma_{10p}(Ease.1Ago)$$
- $$\beta_{2sp} = \gamma_{20p}(Ease.2Ago)$$
- $$\beta_{3sp} = \gamma_{30p}(Ease.3Ago)$$
- $$\beta_{4sp} = \gamma_{40p}(RI.1Ago)$$
- $$\beta_{5sp} = \gamma_{50p}(Retrieval.Number)$$

■ Level 3 (Participant)

- $\gamma_{00p} = \delta_{000} + \delta_{001}(P.Average.Accuracy.Session1) + v_{00p}$
- $\gamma_{10p} = \delta_{100}(Ease.1Ago)$
- $\gamma_{20p} = \delta_{200}(Ease.2Ago)$
- $\gamma_{30p} = \delta_{300}(Ease.3Ago)$
- $\gamma_{40p} = \delta_{400}(RI.1Ago)$
- $\gamma_{50p} = \delta_{500}(Retrieval.Number)$
- $\gamma_{01p} = \delta_{010}(Duration.per.Retrieval)$
- $\gamma_{02p} = \delta_{020}(Schedule.Expansion.Rate)$
- $\gamma_{03p} = \delta_{030}(1^{st}.Retention.Interval)$
- $\gamma_{04p} = \delta_{040}(Judgment.of.Learning)$

This model's total explanatory power was substantial (conditional $R^2 = 0.64$). The model's intercept was 1.37 (95% CI [0.43, 2.30], $t(4594) = 2.86$, $p < .01$). Similar to the lookback window analysis, the ease of the previous three retrievals positively predicted current ease. The participant's initial Judgment of Learning rating was also positively predictive of ease, as was the participant's overall accuracy on Session 1 retrievals (all retrievals except the final, delayed retrieval attempt). The size of the previous retention interval negatively predicted ease, as did the expansion rate of the schedule. The average space between retrieval attempts (*Duration.per.Retrieval*) was notably non-significant. Full results may be found in Table 2 below.

Table 2. All Major Predictors of Ease

<i>Predictors</i>	Ease		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	1.35	0.32 – 2.38	0.010
Ease.1Ago	0.51	0.48 – 0.54	<0.001
Ease.2Ago	0.19	0.15 – 0.22	<0.001
Ease.3Ago	0.08	0.05 – 0.11	<0.001
RI.1Ago	-0.01	-0.01 – -0.00	0.021
Retrieval.Number	0.04	-0.02 – 0.10	0.207
Duration.per.Retrieval	-0.00	-0.02 – 0.01	0.748
Schedule.Expansion.Rate	-0.92	-1.39 – -0.44	<0.001
1 st .Retention.Interval	-0.01	-0.06 – 0.04	0.737
Judgment.of.Learning	0.15	0.08 – 0.22	<0.001
P.Average.Accuracy.Session1	4.03	3.23 – 4.83	<0.001
Random Effects			
σ^2	10.66		
$\tau_{00 \text{ par}}$	0.20		
$\tau_{00 \text{ sch}}$	0.10		
ICC	0.03		
N_{sch}	24		
N_{par}	78		
Observations	4608		
Marginal R^2 / Conditional R^2	0.635 / 0.645		

Mediation: Retention Intervals Driving Ease

Due to the large and highly significant effects of prior ease observed in the preceding analysis, combined with the comparatively smaller and less significant effect of the previous retention interval, we considered the possibility of a mediating effect involving these two variables. A Baron and Kenny mediation analysis was conducted to assess whether the

relationship between *RI.2Ago* and *Ease* was mediated by *Ease.1Ago*. This model is depicted in Figure 15. In Step 1 of the mediation model, *RI.2Ago* significantly predicted *Ease*, $\beta = -0.032$, $t(9671) = -18.84$, $p < .001$. Step 2 showed that *RI.2Ago* significantly predicted *Ease.1Ago*, $\beta = -0.027$, $t(9588) = -15.77$, $p < .001$. Step 3 of the mediation process showed that the mediator (*Ease_1_Ago*) significantly predicted *Ease*, $\beta = 0.666$, $t(8957) = 81.11$, $p < .001$, controlling for *RI.2Ago*. A Sobel test was conducted and found partial mediation in the model ($z = -25.45$, $p < .001$).

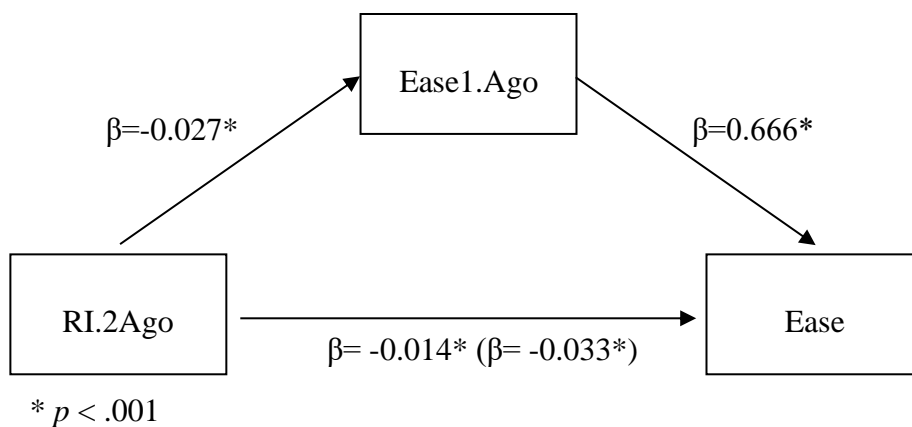


Figure 15. Retention Interval Mediation: *RI.2Ago* Acting Through *Ease.1Ago*

An equivalent model, depicted in Figure 16, explored if a mediating effect might also be occurring between earlier retention intervals and subsequent retrieval difficulties. This next Baron and Kenny mediation analysis was conducted to assess whether the relationship between *RI.3Ago* and *Ease* was mediated by *Ease.2Ago*. In Step 1 of the mediation model, *RI.3Ago* significantly predicted *Ease*, $\beta = -0.035$, $t(7787) = -16.47$, $p < .001$. Step 2 showed that *RI.3Ago* significantly predicted *Ease.2Ago*, $\beta = -0.026$, $t(7674) = -12.54$, $p < .001$. Step 3 of the mediation process showed that the mediator (*Ease.2Ago*) significantly predicted *Ease*, $\beta = 0.571$, $t(7175) = 54.64$, $p < .001$, controlling for *RI.2Ago*. A Sobel test was conducted and found partial mediation in the model ($z = -24.23$, $p < .001$).

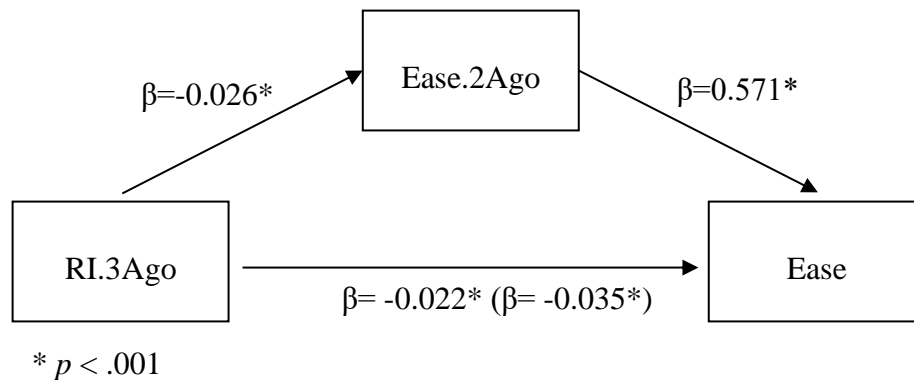


Figure 16. Retention Interval Mediation: *RI.3Ago* Acting Through *Ease.2Ago*

A final Baron and Kenny mediation analysis was conducted to go one step further back and assess whether the relationship between *RI.4Ago* and *Ease* was mediated by *Ease.3Ago*. This is depicted in Figure 17. In Step 1 of the mediation model, *RI.4Ago* significantly predicted *Ease*, $\beta = -0.036$, $t(5918) = -13.24$, $p < .001$. Step 2 showed that *RI.4Ago* significantly predicted *Ease.3Ago*, $\beta = -0.026$, $t(5799) = -9.98$, $p < .001$. Step 3 of the mediation process showed that the mediator (*Ease.3Ago*) significantly predicted *Ease*, $\beta = 0.496$, $t(5435) = 37.81$, $p < .001$, controlling for *RI.4Ago*. A Sobel test was conducted and found partial mediation in the model ($z = -22.05$, $p < .001$).

Predicting Delayed Retrieval Accuracy

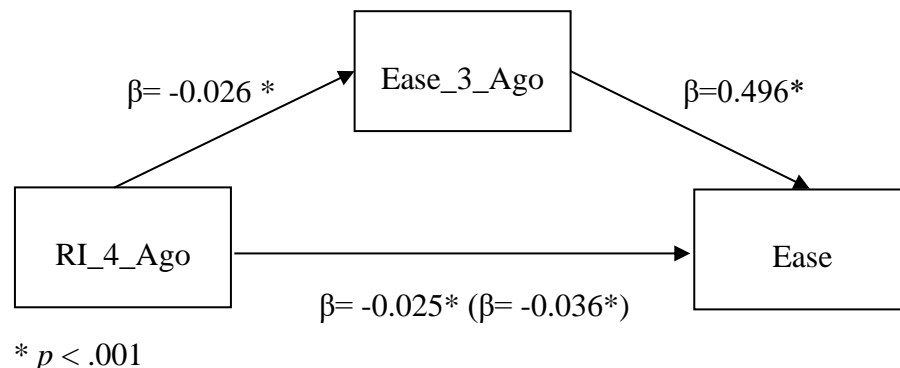


Figure 17. Retention Interval Mediation: *RI.4Ago* Acting Through *Ease.3Ago*

Next, a hierarchical linear model was fit to investigate the aspects and characteristics of schedules that predict long-term retention. Long-term retention was measured through performance on a delayed retrieval attempt occurring three days after the learning period (Session 2 of this experiment). Because we were interested in using characteristics of the schedules to predict performance and the DV in hierarchical models must be a Level 1 variable, we omitted retrieval-level variables and fit the model with only two levels. Level 1 represented the Schedule level and Level 2 represented the Participant level. The same dataset as the previous models was used, except that the data were aggregated across all retrievals in a schedule, per participant. Thus, each schedule generated one data point per participant. The complement of the Levenshtein distance, converted to a percentage, was used as the outcome variable. The Levenshtein distance describes the difference between two strings, as measured by the number of deletions, insertions, or substitutions required to arrive at the correct string (Levenshtein, 1966). If, for instance, the participant responded ‘613’ where the correct response was ‘619’, the Levenshtein distance would be 1, resulting in a complementary Levenshtein percentage of 66.6% (i.e., 2 of 3 digits were correct). For simplicity, we will refer to this value as the Levenshtein percentage (*LevdPct.S2*). The formula for this model is presented below.

- Level1 (Schedule)

- $LevdPct.S2_{sp} = \beta_{0p} + e_{sp}$

- Level 2 (Participant)

- $\beta_{0p} = \gamma_{00} + u_{0p}$

Results from this analysis indicated that 73% of the variability in retrieval *Ease* was at Level 1, the schedule level, ($\sigma_e^2 = 0.11$), and 27% was at Level 2, the participant level, ($\sigma_u^2 = 0.04$). The fully unconditional model thus indicated that there was sufficient variability for

further analyses.

Relative Spacing: Constant, Contracting, Expanding

The next step of our analysis consisted of comparing the efficacy of the primary three categories of schedule types. These primary schedule types were contracting, expanding, and constant (again, these types reflect the change in successive retention intervals over time). These schedules were matched in terms of 1) total schedule duration, 2) number of retrieval attempts, and, therefore, 3) the average retention interval between retrieval attempts. This left the relative spacing of retrieval attempts as the only differentiating factor between the schedule types. In

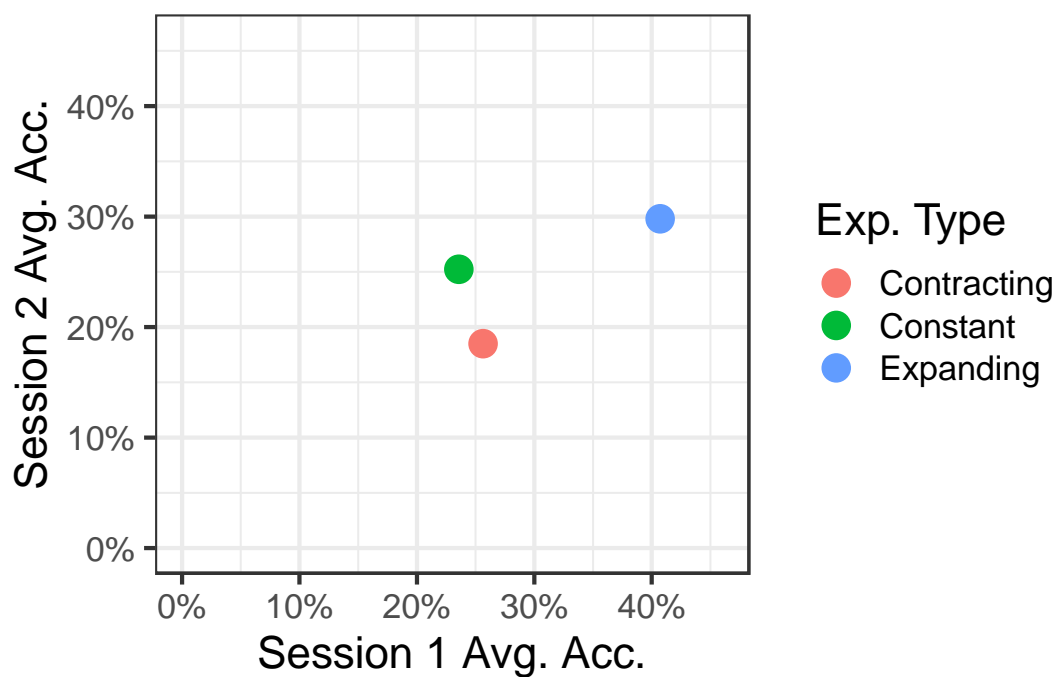


Figure 18. Session 1 and Session 2 Performance, by Expansion Type

Figure 18 we see a scatterplot of the aggregated Session 1 and Session 2 accuracies for each of the three schedule types. Through visual inspection, we see that expanding schedules produced the highest average accuracy in both Sessions (with Session 2 being the single delayed retrieval attempt). Contracting and constant schedules performed similarly in Session 1, though constant schedules performed better in Session 2. In Figure 19 we see a view of the same scatterplot of

performance on Session 1 and Session 2, although broken out to display the average performance of each schedule. There were nine schedules per expansion type. As a final visualization of performance by schedule we plotted the average performance of each schedule over the full

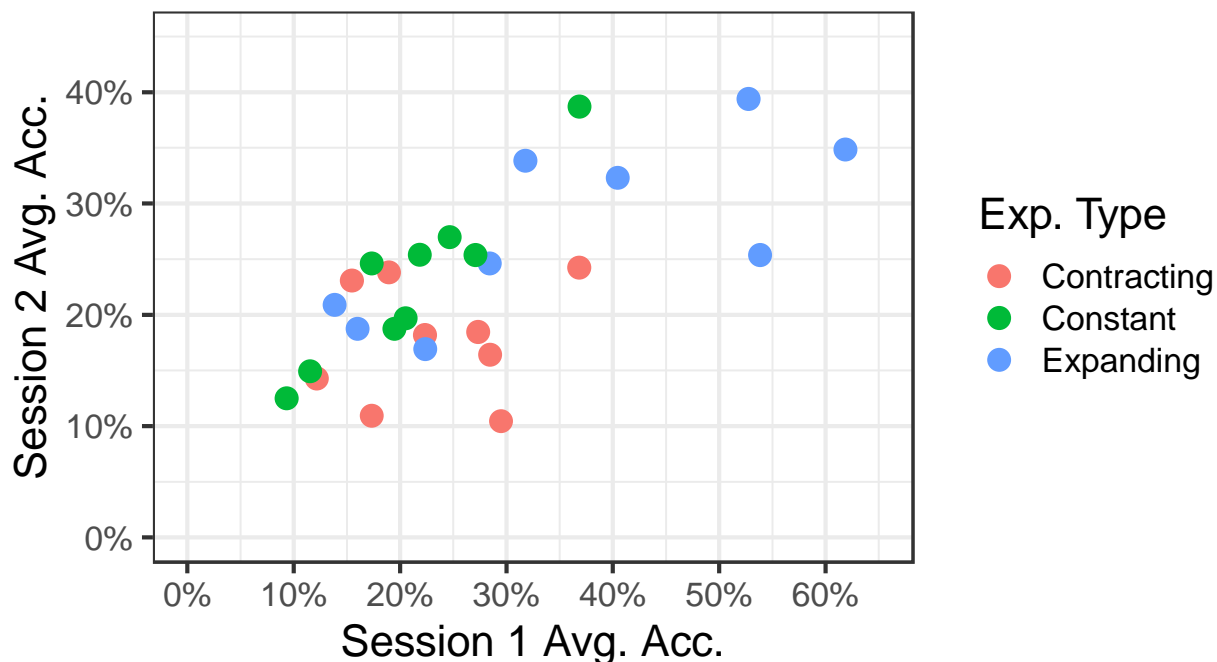


Figure 19. Session 1 and Session 2 Performance, by Schedule

timeframe of Session 1 (nearly 2 hours), for each retrieval attempt. This can be seen in Figure 20. The schedules are separated according to expansion type, leaving nine schedules per grid. Each schedule is colored according to its matched trio, which refers to the group of schedules that have the exact same total schedule duration and number of retrieval attempts. Each trio consists of one constant, one expanding, and one contracting schedule. These trios were derived as detailed in the Experiment 1 Method section (Design).

To further investigate the effects of relative spacing this we fitted a linear mixed model to predict a schedule's Session 2 Levenshtein Percentage ($Sch.LevD.\%Session2$) using only

Expansion.Type as a predictor. The model's total explanatory power was substantial (conditional

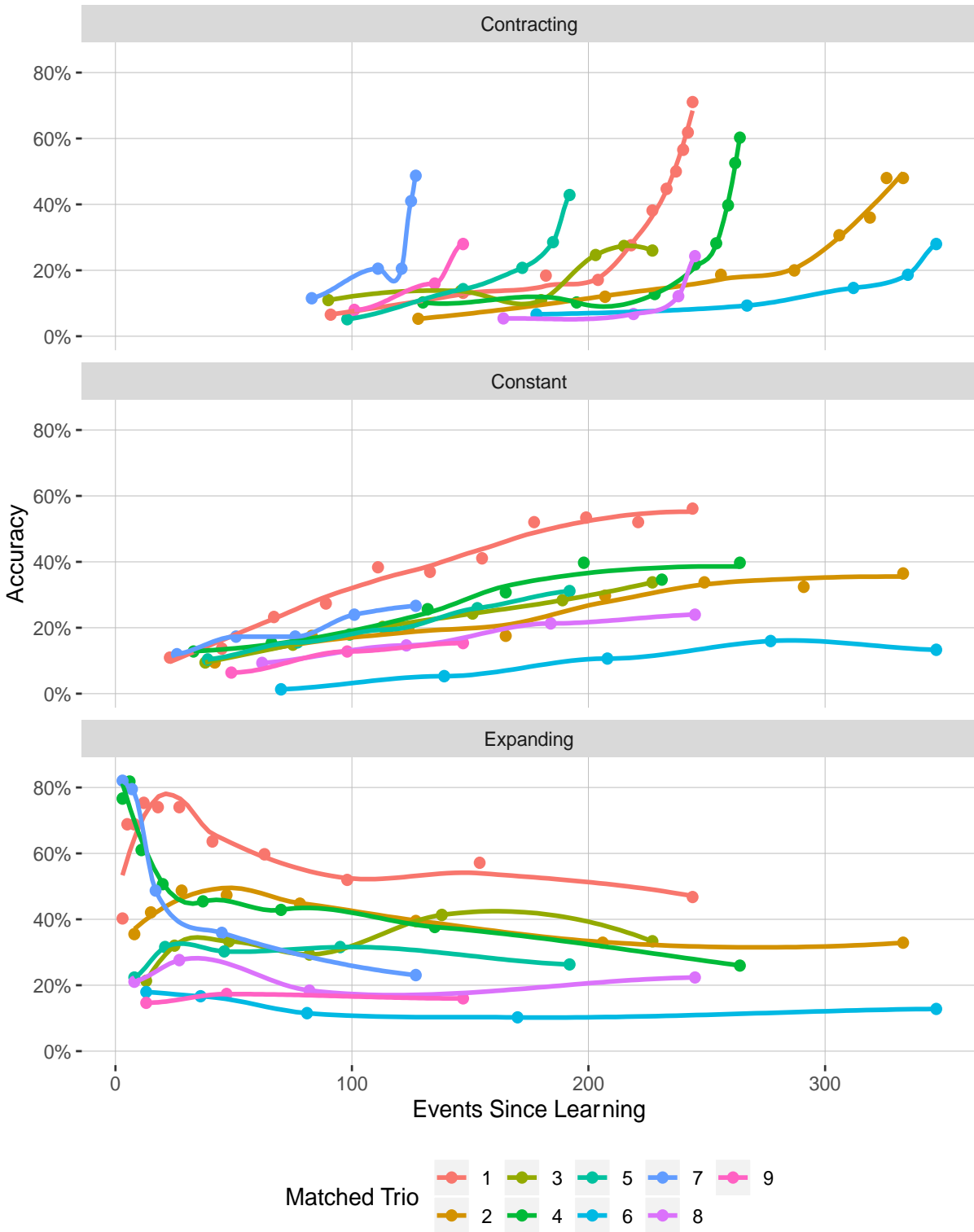


Figure 20. Session Accuracy Over Time

$R^2 = 0.29$). The model's intercept was 0.30 (95% CI [0.25, 0.36], $t(1752) = 10.48$, $p < .001$). Because *Expansion.Type* is a categorical predictor, contracting schedules were used as the baseline. We observed that the effect of *Expansion.Type* [constant] was significantly positive with respect to the contracting schedules (beta = 0.05, 95% CI [0.01, 0.09], $t(1752) = 2.72$, $p < .01$; Std. beta = 0.13, 95% CI [0.04, 0.23]). The effect of *Expansion.Type* [expanding] was also significantly positive with respect to the contracting schedules (beta = 0.10, 95% CI [0.06, 0.14], $t(1752) = 5.06$, $p < .001$; Std. beta = 0.25, 95% CI [0.15, 0.35]).

Post hoc comparisons using the Tukey HSD test indicated that the mean Levenshtein percentage for the constant condition ($M = 0.358$) was significantly greater than the contracting condition ($M = 0.304$), $p < .05$. The mean Levenshtein percentage for the expanding condition ($M = 0.404$) was also significantly greater than the contracting condition ($M = 0.304$), $p < .01$. The mean Levenshtein percentage for the expanding condition ($M = 0.404$) was not significantly greater than the contracting condition ($M = 0.358$), though it approached significance ($p = .053$).

All Major Predictors of Long-Term Retention

Next, we looked at a model containing all relevant predictors. This model is conceptually similar to the other 'All Major Predictors' model described earlier, although here delayed Session 2 performance (as measured by the Levenshtein percentage) was used as the DV rather than the ease of the next retrieval. The formula of this model is given below.

■ Level 1 (Schedule)

$$\begin{aligned}
 - \text{LevdPct}_{sp} = & \beta_{0p} + \beta_{1p}(\text{Duration.per.Retrieval}) + \beta_{2p}(\text{Schedule.Expansion.Rate}) + \\
 & \beta_{3p}(\text{Expansion.Type}) + \beta_{4p}(\text{Schedule.Ease}) + \beta_{5p}(\text{Slope.Ease.x10}) + \\
 & \beta_{6p}(\text{Judgment.of.Learning}) + \beta_{7p}(\text{1}^{\text{st}}.\text{Retention.Interval}) + \beta_{8p}(\text{Schedule.Retrievals}) \\
 & + e_{sp}
 \end{aligned}$$

■ Level 2 (Participant)

- $\beta_{0p} = \gamma_{00} + \gamma_{01}(P.Average.Accuracy.Session1) + u_{0p}$
- $\beta_{1p} = \gamma_{10}(Duration.per.Retrieval)$
- $\beta_{2p} = \gamma_{20}(Schedule.Expansion.Rate)$
- $\beta_{3p} = \gamma_{30}(Expansion.Type)$
- $\beta_{4p} = \gamma_{40}(Schedule.Ease)$
- $\beta_{5p} = \gamma_{50}(Slope.Ease.x10)$
- $\beta_{6p} = \gamma_{60}(Judgment.of.Learning)$
- $\beta_{7p} = \gamma_{70}(1^{st}.Retention.Interval)$
- $\beta_{8p} = \gamma_{80}(Schedule.Retrievals)$

In this linear mixed model we aimed to predict each schedule's Session 2 Levenshtein Percentage (*Sch.LevD%.Session2*). Predictors in the model included *Duration.per.Retrieval*, *Schedule.Expansion.Rate*, *Expansion.Type*, *Schedule.Ease*, *Slope.Ease.x10*, *Judgment.of.Learning*, *1st.Retention.Interval*, *Schedule.Retrievals* and *P.Average.Accuracy.Session1*. The model included random effect error terms at each level. The model's total explanatory power was substantial (conditional $R^2 = 0.51$) and the part related to the fixed effects alone (marginal R^2) was 0.43. The model's intercept was 0.02 (95% CI [-0.14, 0.19], $t(1665) = 0.29$, $p = 0.772$). Although the intercept was not significant, many predictors were highly significant. For instance, the effect of *Duration.per.Retrieval* was significantly positive, as was the effect of *Schedule.Ease*, and a participant's average Session 1 accuracy (*PAverage.Accuracy.Session1*). The effect of *1st.Retention.Interval* was significantly negative. Full results may be found in Table 3.

Table 3. All Major Predictors of Long-Term Retention

<i>Predictors</i>	<i>Sch.LevD.%.Session2</i>		
	<i>Estimates</i>	<i>CI</i>	<i>p</i>
(Intercept)	0.02	-0.14 – 0.19	0.772
<i>Duration.per.Retrieval</i>	0.00	0.00 – 0.01	<0.001
<i>Schedule.Expansion.Rate</i>	-0.04	-0.08 – 0.00	0.071
<i>Expansion.Type [constant]</i>	-0.09	-0.19 – 0.01	0.076
<i>Expansion.Type [expanding]</i>	-0.22	-0.39 – -0.05	0.012
<i>Schedule.Ease</i>	0.05	0.05 – 0.06	<0.001
<i>Slope.Ease.x10</i>	0.02	-0.01 – 0.05	0.116
<i>Judgment.of.Learning</i>	0.01	0.00 – 0.02	0.022
<i>1st.Retention.Interval</i>	-0.02	-0.03 – -0.01	<0.001
<i>Schedule.Retrievals</i>	0.00	-0.01 – 0.01	0.570
<i>P.Average.Accuracy.Session1</i>	0.38	0.22 – 0.54	<0.001
Random Effects			
σ^2	0.08		
$\tau_{00 \text{ par}}$	0.01		
ICC	0.13		
N_{par}	67		
Observations	1678		
Marginal R^2 / Conditional R^2	0.435 / 0.507		

Discussion: Experiment 1

The aim of Experiment 1 was to assess the factors contributing to the successful learning and retrieval of declarative information. In particular, this experiment tested the desirable

difficulty hypothesis more directly than previous works, which often claimed evidence of desirable difficulty in a post-hoc manner. The first step in this investigation, then, was to validate the proposed CASD scale and compare its efficacy to that of response latency, the most common continuous, retrieval-level proxy for retrieval difficulty in previous literature (Benjamin et al., 1998; Gardiner et al., 1973; Karpicke & Roediger, 2007; Pyc & Rawson, 2009). It is important to note that generally when the difficulty of a retrieval is referenced, difficulty it is implied to refer not only to the subjective feeling, but also the probability that a retrieval would have been successful (Kang et al., 2014; Karpicke & Roediger, 2007; Landauer & Bjork, 1978; Roediger & Karpicke, 2006). Given that the probability of successful retrieval of a piece of declarative information is a probabilistic and continuous function, rather than a dichotomous one (Averell & Heathcote, 2011; Wixted & Ebbesen, 1991), it is important to identify the best continuous estimate for the likelihood of retrieval success.

CASD Scale Efficacy

This work found that, although both response latency and the CASD scale were significantly correlated with accuracy (and therefore the probability of retrieval success), the CASD scale explained a full 70% more variance in a schedule's accuracy than did response latency. In this respect the CASD scale provided a better continuous, retrieval-level approximation of the likelihood of retrieval success than other existing methods.

One notable drawback of the CASD scale as compared to response latency is the additional time and effort imposed on the learner by asking him or her to respond to the CASD scale after each retrieval attempt. Despite requiring marginally more time and effort, responding to the CASD scale could, in theory, function in much the same way as other Judgment-of-Learning (JOL) tasks, which are beneficial to the learning process in and of themselves. The

current experimental design was not built to address this possibility, though we hope follow-up research might investigate this potential effect.

Despite the strikingly high variance explained by the CASD scale, Figure 11 clearly illustrates that there is a skew towards both ends of the scale. For incorrect retrievals, this represented feeling far from the correct answer, and for correct retrievals represented feeling that the retrieval was easy. This could be an artefact of the scale being administered after feedback had been provided. It is possible, then, that the “knew-it-all-along” effect and its cousin, the “was-never-going-to-remember-it” effect were at work here. This does not rescind the explanatory value of the scale in any way, although it does indicate there may be room for refinement. For instance, it is possible that the CASD scale was over-specified, seeing as 2 of the 14 responses constituted the great majority of sampled values. It is possible that a modified scale could explain a similar amount of variance (or more) with fewer response options, while prompting a more normal response distribution.

Predicting Ease

Having demonstrated the efficacy of the CASD scale, we set out to utilize Ease as our dependent variable in 3-level hierarchical linear models. The null model showed that 71% of the variability in Ease resided at the retrieval level (Level 1), 10% at the schedule level (Level 2), and 19% at the participant level (Level 3). This 3-level structure explained significantly more variance than the 2-level equivalent. These findings demonstrate that a disproportionate percent of variance is explained at the level of the retrieval (Level 1), and also that a notable degree of variation exists at the person level (Level 3). It is apparent that learners vary a great deal in their aptitudes, effort, or both, further reinforcing the importance of including variance at the participant level. An ideal schedule for one learner may be ill-suited to another, even if they have

the same learning goals. This makes it all the more important to document these inter-individual differences and optimize schedules accordingly.

Linear Vs. Quadratic Ease Model

With the hierarchical structure justified, we next investigated one of the core tenets of the desirable difficulty hypothesis, namely that it is possible for retrievals to be too easy, compromising long-term retention. To test this, we compared two hierarchical models where the previous retrieval ease (per the CASD scale) was used to predict current ease. In one model the previous ease term was quadratic, and in the other it was linear. The model with the linear previous ease term explained significantly more variance than the model with the quadratic previous ease term. This is evidence against the notion that a retrieval can be too easy. The fact that the linear model was superior indicates that there was not a notable inflection point after which a higher ease of a previous retrieval began to negatively predict current ease. Instead, the observed relationship was that the *easier* the previous retrieval was, the *easier* the current retrieval was likely to be, with no qualifications or upper bound.

Lookback Window

Continuing this exploration, we constructed another hierarchical model, described here as a lookback window, designed to test the size and directionality of the main effects. The difficulties of the four previous retrievals were used as predictors. The desirable difficulty hypothesis would predict that incurring additional difficulty at some Time(N) would result in decreased difficulty at some point in the future, Time(N+X) (Bjork, 1994). Put another way, per the desirable difficulty hypothesis you would expect an inverse relationship between difficulty at some Time(N) and Time(N+X), at least for certain difficulties and certain intervals between retrieval attempts. This core tenet of the desirable difficulty effect was not supported by the

findings. Instead, findings showed that the ease of each of the four previous retrievals independently and positively predicted the ease of the current retrieval. Similar to the finding of the linear versus quadratic comparison, we found that the *easier* the previous four retrievals were, the *easier* the current retrieval was likely to be. Notably, though perhaps expectedly, the effect sizes decreased as the temporal distance between a prior retrieval and the present retrieval increased. That is to say, each retrieval had more influence on the current retrieval's ease than the retrieval preceding it, with all effects in the positive direction as depicted in Figure 14.

All Major Predictors of Ease

To this point, we have only looked at prior ease as a predictor. In this next model, then, we looked to include all relevant predictors of present retrieval difficulty. This was done to produce as comprehensive of a model as possible to concurrently assess the contributions of many factors hypothesized to influence difficulty, as well as observe their relative effect sizes and directions. The predictors entered included:

- Level 1 (Retrieval): *Ease.1Ago*, *Ease.2Ago*, *Ease.3Ago*, *Retention.Interval.1Ago*, and *Retrieval.Number*.
- Level 2, (Schedule): *Duration.per.Retrieval*, *Schedule.Expansion.Rate*, *1st.Retention.Interval*, and *Judgment.of.Learning*.
- Level 3 (Participant): *P.Average.Accuracy.Session1*.

This model demonstrated that, similar to the lookback window, each of the three prior retrieval eases were uniquely and positively predictive of the current retrieval ease, with p-values approaching zero. As explained earlier, this is evidence against the desirable difficulty hypothesis, which would have predicted at least one effect in the negative direction. Continuing with the retrieval level factors (Level 1), *Retention.Interval.1Ago* (the space between the

previous retrieval and the present one) was negatively predictive of current retrieval ease, though only reaching significance at the $p=.05$ threshold. The *Retrieval.Number*, or how many retrievals the participant had attempted for that particular schedule, was not significant. This is perhaps counter to the expectation that the greater the number of retrieval attempts, the easier the present retrieval would tend to be.

At the schedule level (Level 2), the *Duration.per.Retrieval* (the total schedule duration divided by the number of retrievals) was not significant. This is notable, as intuition might lead one to expect that the longer the average retention interval the more difficult each retrieval would be. We suspect that other variables entered into the model may be accounting for this variance. The *Schedule.Expansion.Rate* was significant and negatively predictive of retrieval ease, indicating that the slower the expansion, the easier the next retrieval tended to be. Next, the *Judgment.of.Learning* rating that participants made after the first learning event was found to be positively predictive of current ease. That is, participants' initial confidence regarding how well they learned the prompt and response pair was accurate to a significant degree.

Rounding out the Level 2 predictors, the size of the *1st.Retention.Interval* was also found to be non-significant. This does not fit with previous work that found the size of the first retention interval to not only be positively predictive of current ease, but to be the major determinant of a schedule's long-term retention efficacy (Karpicke & Roediger, 2007).

Finally, at the participant level (Level 3), participants' average Session 1 accuracy was significantly and positively predictive of current retrieval ease. This finding further confirms the presence of inter-individual differences and the importance of accounting for participant-level variables.

In addition to some of the surprising or otherwise notable findings described above, the

relatively small influence of the previous retention interval (*Retention.Interval.1Ago*) seemed unusual to the authors, especially when compared to the highly significant, larger effects of the ease of prior retrievals. From a theoretical standpoint, one might expect the size of the previous retention interval to be one of the largest determinants of future retrieval likelihood (and therefore, difficulty). This led us to assess whether there might be some mediating effects at work.

Mediation: Retention Intervals Driving Ease

For the reasons elucidated above, we explored whether the retention interval before a retrieval might be influencing the difficulty of that retrieval, which in turn was influencing the present ease. Put another way, we looked into whether the ease of an earlier retrieval was mediating the relationship between a preceding retention interval and the ease of the present retrieval. This is best illustrated in the mediation diagrams in the Results section of Experiment 1 (Figures 15, 16, and 17).

The first mediation model involved the *Retention.Interval.2Ago* predicting *Ease.1Ago*, which in turn predicted the present Ease (where *Ease.1Ago* was the mediator). Partial mediation was found to be occurring, demonstrating that some of the variance that might have been explained by the *Retention.Interval.2Ago* was acting through *Ease.1Ago*, which was in turn largely predicting the present retrieval Ease. This mediating effect may go a long way in explaining the unexpectedly low predictive power (and lower significance) of previous retention intervals and the much larger (and higher significance) predictive power of the eases of previous retrievals.

We also looked for -and found- equivalent mediating effects for earlier retrievals. For instance, we showed that *Retention.Interval.3Ago* was acting through *Ease.2Ago* (the mediator),

which was in turn predicting the present retrieval Ease. Going even further back, we also showed that *Retention.Interval.4Ago* was acting through *Ease.3Ago* (the mediator), which was in turn predicting the present retrieval Ease.

The identification of these mediating effects is especially relevant in light of the finding reported earlier that each of *Ease.1Ago*, *Ease.2Ago*, and *Ease.3Ago* were shown to be uniquely predictive of the present retrieval ease. In conjunction with these mediating effects, then, we have shown that the retention interval preceding each of these three retrieval attempts was responsible for a significant portion of the variance in present retrieval ease, even if this variance *appeared* to be almost entirely attributable to the ease of previous retrievals. To boil these findings down to their simplest form, prior retention intervals were shown to drive the difficulty of the next retrieval, and the difficulty of that retrieval in turn largely drove the difficulty of the next one.

Predicting Delayed Retrieval Accuracy

To this point we have primarily concerned ourselves with the prediction of the accuracy and ease of next retrieval. That is, *Ease* has been used as the Dependent Variable in the preceding models. It is the author's view that this is the most relevant approach to studying declarative learning schedules, treating each retrieval attempt as simply the most recent component of a growing series of retrieval attempts that together constitute a learning schedule. Note that in this view no distinction is made between a "learning" period, and some "delayed testing" period. Instead, every retrieval is assumed to be part of an ever-growing learning period with an evolving availability and forgetting pattern. This interpretation is preferred because we know retrieval attempts do indeed produce learning (rather than simply revealing previous learning) and making a distinction between "learning" and "testing" periods thus feels arbitrary.

In many other works, however, a different approach is taken where the primary DV of interest is the ability of a learning schedule to produce successful long-term retention. This is a schedule-level DV, one level higher than the retrieval-level DV used in the previously described three-level models. In the interest of comparing the present results with the larger body of learning literature the authors elected to also investigate schedule-level measures predictive of long-term retention success.

To do this we created a two-level hierarchical model with success on the final retrieval attempt (Session 2, occurring 3 days after Session 1) serving as the DV. Success was operationalized as the Levenshtein percentage. Again, this Levenshtein percentage refers to the number of deletions, insertions, and substitutions required to go from the given response to the correct response. This model was two levels rather than three because the DV must be a Level 1 variable in hierarchical linear models. The prior dataset was therefore aggregated by participant-schedule to result in one data point per schedule, per participant. Level 1 of this model referred to the schedule level, and Level 2 referred to the participant level.

Relative Spacing: Expanding, Contracting, Constant

First on the list of schedule-level analyses was an exploration of the effect of relative spacing. As described earlier there are mixed findings regarding relative spacing, including whether expanding or constant schedules are superior (Bjork, 1988; Camp, 1996; Kang et al., 2014; Karpicke & Roediger, 2007; Logan & Balota, 2008), and even whether relative spacing produces any effect at all (Karpicke & Bauernschmidt, 2011). To assess the effect of a schedule's expansion type (contracting, constant, or expanding) on long-term retention, expansion type was input as a predictor in the previously described 2-level model. This model, in conjunction with post-hoc analyses, revealed that constant and expanding schedules produced

greater long-term retention than contracting schedules. This is strong evidence that relative spacing does, indeed, have an effect on long-term retention. We also found that expanding schedules produced a higher average success rate on the long-term retention test than did constant schedules, although this effect was just shy of statistical significance at the $p=.05$ threshold ($p=.053$).

All Major Predictors of Long-Term Retention

In the final analysis of Experiment 1 we created a hierarchical model with all relevant predictors of delayed retrieval accuracy. The Levenshtein distance was used as the DV in this model, and the two-level hierarchical structure, as in the previous two-level model, defined Level 1 as the schedule, and Level 2 as the participant level. There are a number of notable findings from this model. For instance, the *Duration.per.Retrieval*, which was non-significant in the preceding three-level model, was now significantly positive. This means that the greater the average interval between retrieval attempts, the better participants were likely to do on the delayed retrieval attempt. This appears to be clear-cut evidence of the spacing effect. The average *Schedule.Ease* was also significant and positively predictive of final retrieval accuracy, corroborating similar findings reported previously.

Diverging from the findings of the three-level model predicting current *Ease*, however, expanding schedules were found to be less effective than contracting schedules, after controlling for *Schedule.Ease*. While this may seem counterintuitive, this reversal of effect direction only appeared when 1) controlling for *Schedule.Ease*, and 2) predicting long-term retention success, rather than predicting *Ease* at the retrieval level. This is of critical importance, because contracting schedules, by definition, have retrievals more and more densely packed together as the schedule progresses, whereas expanding schedules have the opposite pattern. The effect this

has on retrieval ease over time is plainly seen in Figure 20. It is likely, then, that the last retrievals in a contracting schedule will be easier than the last retrievals in an expanding schedule (especially after controlling for *Schedule.Ease*). And as the lookback window earlier demonstrated, the closer in temporal proximity, the larger the influence of that retrieval's ease will be. This gives contracting schedules an advantage in this sense, as their easier retrievals were likely to be more recent. Again, this only occurs when *Schedule.Ease* is controlled. When it is not, as reported in the Relative Spacing section where relative expansion rates were directly compared, expanding intervals produced greater long-term retention than did contracting schedules.

The slope of the retrieval ease (*Slope.Ease*) was not significant, indicating that the trajectory of difficulty did not play a role in long-term retention, controlling for all the other variables in the model. Likewise, *Schedule.Retrievals*, the total number of retrievals in the schedule, was not significant. This is interesting, as one might expect more retrieval attempts to result in better long-term retention. Similar to the earlier findings of the three-level analysis, the *Judgment.of.Learning* rating was significant and positively predictive of long-term retention.

The *1st.Retention.Interval* was significant and negatively predictive of long-term retention. This means that the shorter the first retention interval was, the less likely a schedule was to be remembered after a sizeable delay. This is directly counter to previous research claiming that the first retention interval was in fact the dominant predictor of the long-term efficacy of a learning schedule (Karpicke & Roediger, 2007). If interpreted per the argument in the original work, this is another point against the desirable difficulty hypothesis, or at least not in its favor. This is because the original work posited that the *increased* first retention interval led to *increased* difficulty, which then led to greater long-term retention efficacy. Instead, this

work found the exact opposite: that a *decreased* first retention interval led to *decreased* difficulty, which then led to greater long-term retention efficacy. This fits with the previously described findings in demonstrating a notable lack of evidence in favor of the desirable difficulty hypothesis, after controlling for other relevant variables. Put another way, all signs point to the fact that *ease* is desirable – not difficulty –, and there appears to be no difficulty that is inherently beneficial to the learner.

Experiment 2

In this experiment we aimed to adaptively target a specific retrieval difficulty by using the difficulty of the previous retrievals to determine the spacing after which the next retrieval should occur. This concept is well illustrated by the mediation effects described in Experiment 1, wherein the previous retention interval is shown to predict the current difficulty. This experiment aims to act as a proof-of-concept for the ability to adaptively tailor learning schedules to meet specific availability aims. Because such an adaptive scheduling algorithm accounts for previous difficulty and adjusts itself accordingly, it may thus automatically account for the complexity of the to-be-learned information as well as the ability of the learner. We thus aim to address a major drawback of existing research into declarative learning schedules, where stimulus complexity is almost entirely ignored. Implicit in such works is the assumption that the difficulty of the to-be-learned information is irrelevant and does not influence the “ideal” learning schedule. This does not seem to be a reasonable assumption, as it seems odd to assume that a prompt-response pair like “2+2=4” would be best learned via the same schedule, with the same space between retrieval attempts, as something like “ $x = (-b \pm \sqrt{(b^2-4ac)}) / (2a)$ ”. This work was in part designed to account for the varying levels of stimulus difficulty.

Method

Participants

Sixty-five undergraduate students at least 18 years of age enrolled at a large university in North Carolina participated, and were compensated with course credit.

Materials

This study was administered remotely and was entirely online. Participants completed the study on a computer of their choice (not provided) at a time of their choosing. The experimentation platform Psytoolkit was used to build and subsequently host the experiment (Stoet, 2010; Stoet, 2017).

Design

Experiment 2 was similar to Experiment 1, although only eight schedules were compared. This study employed a partially asymmetric 2x2 factorial design. These factors refer to stimuli difficulty (simple vs. complex), and schedule type (fixed vs. adaptive).

For the less complex stimuli a randomly generated 3x3 black and white grid served as the prompt, paired with a 3-digit number as the response. For the more complex stimuli a randomly generated 4x4 black and white grid served as the prompt, paired with a 4-digit number as the response.

The next manipulated factor was schedule type. Of the eight total schedules, four were adaptive, where an algorithm was used to estimate how long to delay the next retention interval to elicit a target retrieval difficulty, based on prior retrieval difficulties. Of these four adaptive schedules, two targeted high accuracy and two targeted moderate accuracy. The high accuracy adaptive schedules targeted an accuracy of 92%, and the moderate accuracy adaptive schedules targeted an accuracy of 61%. These accuracy values correspond with a 13 and 9 on the CASD

scale, respectively, with 1 being the most difficult and 14 being the easiest. Note that both target difficulties were on the side of the correct responses on the CASD scale.

In addition to these four adaptive schedules there were also four fixed schedules, all of which were expanding. Expanding schedules were selected as the comparators because they are the schedule type hypothesized to be best suited to maintaining a constant retrieval difficulty. This is because after each learning event the memory becomes accessible for longer and longer periods of time. Put another way, as the number of times an item is practiced increases, so does the amount of time it takes that memory to decay to some set level of availability (Averell & Heathcote, 2011; Wixted & Ebbesen, 1991). An expanding schedule, then, is the only way to maintain a constant level of difficulty for each retrieval (in theory, at least). Of these expanding schedules, two expanded at a doubling rate and two expanded at a tripling rate. In total this resulted in eight schedules to be compared, described in Table 4.

Table 4. Schedule Construction, Experiment 2

1	Fixed	Simple stimuli (3x3)	Doubling Expansion
2	Fixed	Complex stimuli (4x4)	Doubling Expansion
3	Fixed	Simple stimuli (3x3)	Tripling Expansion
4	Fixed	Complex stimuli (4x4)	Tripling Expansion
5	Adaptive	Simple stimuli (3x3)	Lower Target Ease
6	Adaptive	Simple stimuli (3x3)	Higher Target Ease
7	Adaptive	Complex stimuli (4x4)	Lower Target Ease
8	Adaptive	Complex stimuli (4x4)	Higher Target Ease

Procedure

Participants accessed the experiment by following the provided URL after reading a description of the study on the university's experimentation site, SONA. They completed the experiment using an internet browser, on a computer of their choice. The entire experiment took just under 2 hours, including instructions and practice trials.

The structure of this experiment was largely identical to Experiment 1, described above, from the perspective of the participant. After receiving instruction and performing practice trials, participants learned and attempted to retrieve paired associates as in Experiment 1, though in this study there were only eight schedules. As in Experiment 1, there were three distinct types of events that occurred: initial learning events, retrieval attempts (including feedback and making CASD ratings), and working memory events. See Experiment 1 for a review of these events and their structure. Experiment 2 did not require follow-up or delayed testing, so at the conclusion of the experiment participants were thanked for their time and provided course credit.

Results

In Figure 21, below, we see a scatterplot of each participant's average accuracy on the adaptive schedules as compared to the fixed schedules. The data of three participants were excluded from analyses due to a combination of random responding, extremely low response accuracy, and visual data inspection. It was especially important to exclude these data because the adaptive schedules may behave strangely in cases of non-responding or random responding. This is because the algorithm attempts to shrink the expansion rate after each incorrect response, and if no correct answers come to reverse this trend a single item can be presented to the participant more and more frequently, sometimes dozens of times. It was therefore important to exclude such heavily weighted outliers.

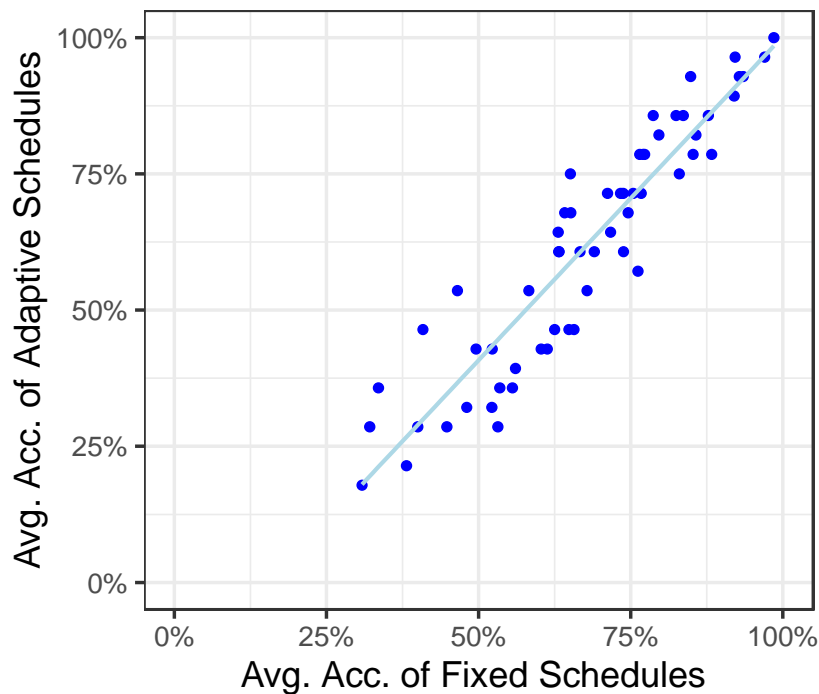


Figure 21: Session 1 Accuracy of Fixed vs. Adaptive Schedules, per Participant

Efficacy of Ease Targeting in Adaptive Schedules

To investigate the effect of a schedule's target ease a One-Way ANOVA was conducted. The main effect of *Target.Ease* was significant: $F(5, 4229) = 35.65, p < .001$; η^2 (partial) = 0.04, 90% CI [0.03, 0.05]. Post hoc comparisons using the Tukey HSD test indicated that the average ease for the adaptive schedules targeting an ease of 9 ($M = 9.12$), the low ease target condition, was closer to an ease of 9 than the adaptive schedules targeting an ease of 13 ($M=10.57$), the high ease target condition, $p < .01$. The low ease condition was also closer to an ease of 9 than the fixed schedule 3 ($M=8.11$), which employed a tripling rate of expansion with the less complex stimuli, $p < .05$. The adaptive schedules targeting an ease of 9 were not significantly closer to an average ease of 9 than fixed schedules 1 (doubling, less complex stimuli), 2 (doubling, more complex stimuli), or 4 (tripling, more complex stimuli), $p > .05$.

An additional post hoc comparison using the Tukey HSD test indicated that the average

ease for the adaptive schedules targeting an ease of 13, the high ease condition, ($M = 10.57$) was closer to an ease of 13 than the adaptive schedules targeting an ease of 9 ($M=9.12$), the low ease target conditions, $p<.01$. The high ease condition was also closer to an ease of 13 than all of the fixed schedules, $p<.05$. These included fixed schedules 1 (doubling, easy stimuli), 2 (doubling, hard stimuli), 3 (tripling, easy stimuli), and 4 (tripling, hard stimuli). Figure 22 depicts linear regression lines for the adaptive schedules, in teal, and the fixed schedules, in red. Only participants with an average accuracy above 50% were included in Figure 22 to avoid the regression lines being too heavily skewed towards the lowest performers. This is because as accuracy decreases, the more closely together the algorithm packs retrievals, biasing the regression line towards these lowest performers which necessarily also have the most retrieval attempts.

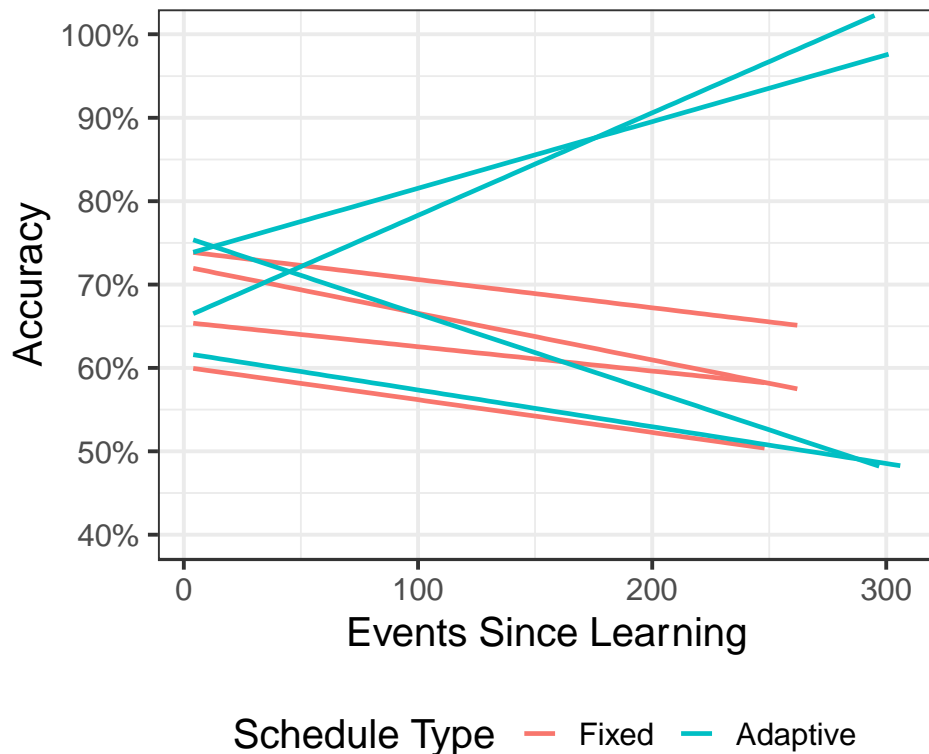


Figure 22: Schedule Accuracy Over Time, Colored by Schedule Type

Comparison of Fixed and Adaptive Schedules

In Figure 23 we have a boxplot visualization of the performance of each of the eight schedules. The four fixed schedules are on the left, and the four adaptive schedules are on the right. The black dots represent the schedule's mean accuracy. Boxes outlined in green correspond to simple stimuli pairs (3x3 grid and 3-digit response), while those outlined in red correspond to complex stimuli pairs (3x3 grid and 3-digit response).

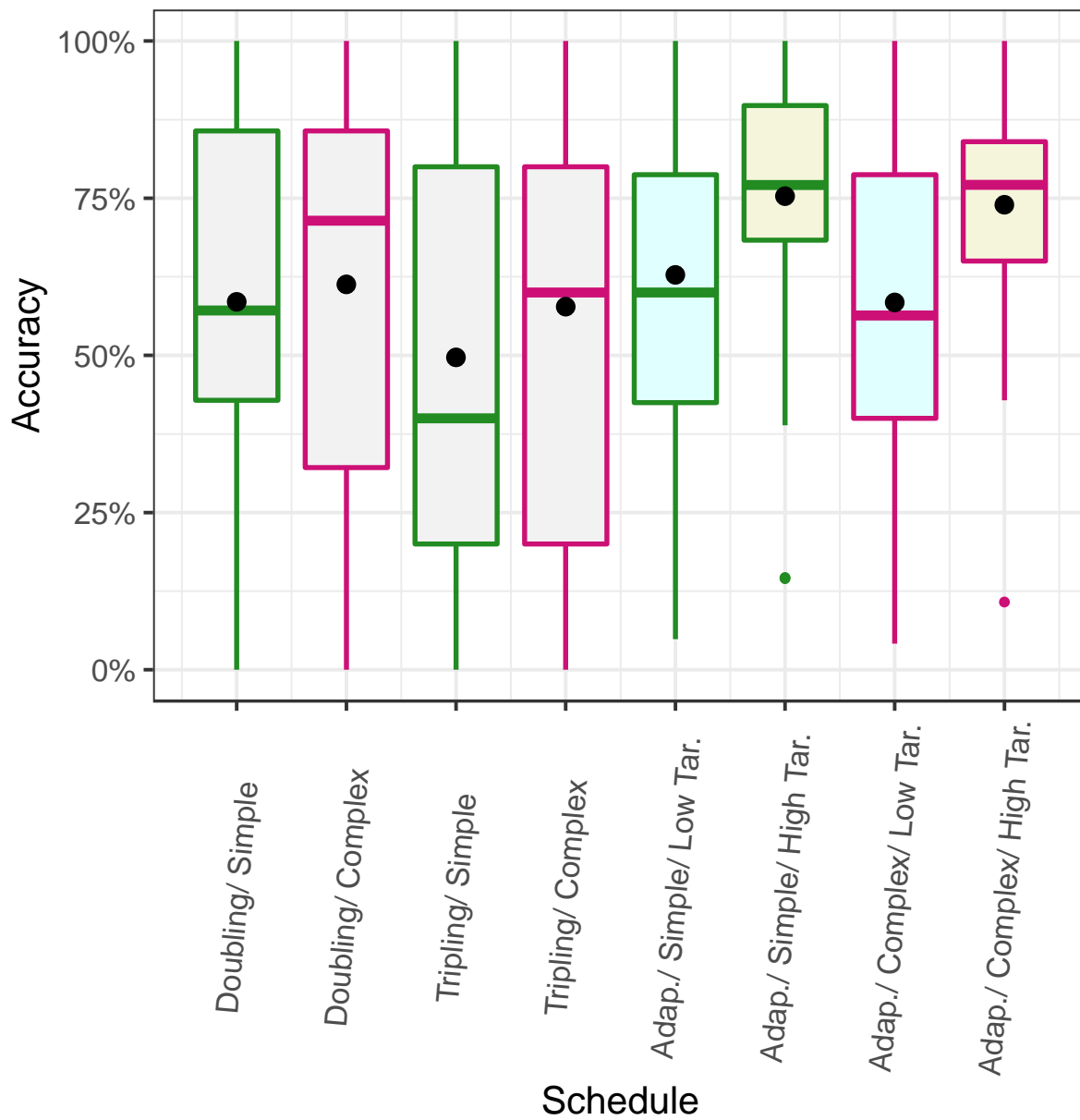


Figure 23: Boxplot of Schedule Performance

correspond to complex stimuli pairs (4x4 grid and 4-digit response). Fill color, used only for the adaptive schedules, indicates the two target accuracies (blue = lower target, yellow = higher target).

Lastly, Levene's test for the homogeneity of variance was conducted to compare the variance of the fixed and adaptive schedules. This test showed that there was indeed less variance in the adaptive schedules than there was for the fixed schedules ($F = 1$, $p = .015$), indicating that the adaptive schedules resulted in a tighter distribution of ease values.

Discussion: Experiment 2

Experiment 2 demonstrated the potential to adaptively build and modify schedules in real time. The goal was to create schedules that are more tailored to 1) the individual, 2) the to-be-learned item, and 3) the goals of the learner, as defined by an availability function. The algorithm employed was relatively simple. It used the target difficulty for each retrieval (which was pre-established), the retrieval's ease, as well as the current expansion rate (how the most recent retention interval compared to the one prior), and then integrated these factors to determine how long the next retention interval should be. For instance, if the target ease was X , the previous retrieval ease was $X-1$ (harder than desired), and the previous expansion rate was Z , then the algorithm might decrease the next expansion rate to $(.9)Z$, as an expansion rate of Z resulted in greater difficulty than intended. Note that the target difficulty in this study was held constant for each retrieval. This was only for proof-of-concept purposes, and in theory the target difficulty (and corresponding availability curve) could certainly be non-linear and vary over time.

This algorithm was largely effective in targeting a specific retrieval ease, both when compared to fixed schedules and when compared to the other adaptive schedules targeting different retrieval difficulties. This is best evidenced by the fact that the adaptive schedules

targeting high ease were significantly closer to their target goal than any of the other schedules assessed. The adaptive schedules targeting moderate ease were “only” significantly closer to their target than one of the four fixed schedules, although they were closer to their target than the other adaptive schedules which were targeting high eases. Taken together, this is strong evidence that the adaptive schedules did indeed result in schedules with an average retrieval ease very close to their predetermined targeted ease.

Levene’s test of homogeneity also revealed that there was less variance in retrieval ease for the adaptive schedules than the fixed schedules. This shows that the adaptive schedules produced more consistent patterns of retrieval ease, with a tighter distribution, than did fixed schedules. This is not entirely surprising given that fixed schedules have no means to correct course once they have begun, and therefore have more potential to “drift” from the expected ease pattern. The adaptive schedules, conversely, play an active and supportive role, decreasing the expansion rate when retrievals are harder than the target and increasing the expansion rate when retrievals are easier than the target. In this manner, the adaptive schedules were successfully able to constrict variance and achieve average schedule eases that were closer to their target eases than the fixed, predetermined schedules.

GENERAL DISCUSSION

The primary aims of this work were tripartite. The foremost aim of Experiment 1 was to assess the factors contributing to the learning and retrieval of declarative information and their relation to one another, with a secondary goal of further validating and utilizing the CASD scale. In particular, Experiment 1 was designed to test the desirable difficulty hypothesis more directly than previous works. The aim of Experiment 2 was to leverage the insights gained in Experiment 1 and employ them to adaptively target specific retrieval difficulties, thereby more intentionally

and directly manipulating the availability curve of the to-be-learned information.

Results from Experiment 1 demonstrated that the CASD scale serves as a far better continuous, retrieval-level approximation of the odds of retrieval success than response latency, explaining a full 50% more variance in retrieval likelihood than response latency. This is despite response latency being the most commonly employed continuous, retrieval-level proxy for difficulty in the literature. The CASD thus scale provides a more nuanced and accurate understanding of retrieval difficulty than previous works and might be considered a methodological contribution to the literature.

In addition to demonstrating the value of the CASD scale, Experiment 1 also broadly observed a lack of evidence in favor of the desirable difficulty framework, at least as traditionally conceptualized. In fact, all gathered evidence suggests that there is no such thing as difficulty that is *inherently* beneficial to the learner. It appears that the optimal level of difficulty is as easy as possible, without exception.

This claim is not made lightly and draws from multiple findings. Among this evidence is the finding that a linear modeling of ease explained more variance than a quadratic equivalent. This showed that the easier a retrieval was, the easier the following retrieval was likely to be, and that there does not appear to be an upper bound on ease after which it becomes undesirable. This is the first point of evidence against the desirable difficulty hypothesis. The next point against the desirable difficulty hypothesis was seen in the lookback window analysis. This model showed that each of the previous four retrieval eases were uniquely and positively predictive of present ease. This is precisely counter the desirable difficulty hypothesis, which states that increased difficulty at Time(N) should lead to decreased difficulty at some point in the future, Time(N+X). This pattern was not observed, as all effects were in the positive direction, again suggesting that

the easier a retrieval is, with no upper bound, the easier the following retrieval is likely to be.

The logical follow-up to this evidence against the desirable difficulty hypothesis, then, is to ask why so many prior works seemed to find evidence of this effect. As alluded, a great deal of this can likely be explained by the post-hoc nature of the majority of the prior works, claiming desirable difficulty to explain an effect after the fact. Such analyses make it difficult to differentiate between other potential explanatory variables. For instance, such retrospective claims do not generally allow for experimental differentiation between the spacing effect and desirable difficulty. This is because increased spacing closely correlates with increased difficulty in the short term and decreased difficulty in the long term. It may then seem that it was the *difficulty* of these earlier retrievals that was driving the long-term benefits of such schedules, rather than the spacing itself. This work allowed for a delineation between these effects, notably finding that after accounting for these other variables –especially total schedule spacing– all evidence in favor of desirable difficulty evaporated.

This is perhaps most plainly exemplified by looking at the effects of *Duration.per.Retrieval* and *Schedule.Ease* in the two-level model predicting long-term retention. Both effects were uniquely and positively predictive of long-term retention (both with $p < .001$). This indicates that you want to maximize not only the total spacing of the learning schedule (equivalent to *Duration.per.Retrieval*), but also the ease of the schedule. The fact that both of these effects were so significant and were able to explain unique variance indicates that 1) they are indeed distinct variables, 2) that certain schedules can produce a divergence between them, and 3) schedules that maximize *both* result in the greatest long-term retention performance. Rephrased, it appears that you want to maximize a schedule's total duration to take advantage of the spacing effect, but want to do this in a way that minimizes the difficulty of the retrievals.

This is further echoed by the exploration of relative spacing, which showed that expanding and constant schedules produced greater long-term retention performance than did contracting schedules. These expanding, contracting, and constant schedules were, however, perfectly matched for total schedule duration, as well as the average retention interval between retrieval attempts. This only leaves difference in expansion rate and retrieval difficulty to explain the success of certain expansion types (expanding, contracting, constant) as compared to others. Expansion rate was shown to be non-significant, while schedule ease was highly significant (and in the positive direction). The desirable difficulty hypothesis has no plausible mechanism to explain the observed differences in long-term retention performance between expanding, contracting, and constant schedules. The finding that the schedule types which resulted in *easier* retrievals were *more* likely to be remembered later is fundamentally at odds with the notion of desirable difficulty.

Instead, the desirable difficulty theory suggests that increasing *Duration.per.Retrieval* (or schedule duration) increases difficulty, and it is this increase in difficulty which creates the benefits to long-term retention. Given the same schedule duration, however, we instead see that *decreasing* difficulty is beneficial for long-term retention, directly counter to this claim. At no point do we see evidence that increasing difficulty has any benefits for the learner, and in fact, it appears that increasing difficulty actively hinders the efficient and successful learning of declarative information. It is the opinion of the authors that a large portion of evidence in favor of the desirable hypothesis is, instead, better interpreted as evidence of the spacing effect. This is to say that the duration of the learning schedule is the primary factor driving long-term retention, despite the fact that difficulty also tends to increase as the schedule duration increases. Per this interpretation, difficulty is an incidental, unintended, and harmful byproduct of increasing the

duration of a learning schedule. In this sense, the desirable difficulty effect may be considered illusory, disappearing when other relevant variables (especially schedule duration) are controlled for.

In addition to largely refuting the desirable difficulty hypothesis, Experiment 1 also served to define the factors beyond difficulty that predicted retrieval success. Specifically, the ease of the previous few retrievals, a schedule's initial Judgment of Learning rating, and the participant's average Session 1 retrieval accuracy all positively predicted future retrieval ease. Conversely, larger prior retention intervals negatively predicted future retrieval ease. When looking at predictors long-term retention, we observed that the size of the first retention interval was negatively predictive of delayed retrieval success. This finding regarding the first retention interval is directly counter to previous findings claiming the effect to be in the positive direction and to be one of the primary determinants of a schedule's long-term retention efficacy (Karpicke & Roediger, 2007). As with other studies, however, these difficulty claims were made after the fact, positing that increasing the first retention interval increased the difficulty of the first retrieval, which in turn decreased difficulty later on. After controlling for other relevant variables, we found strong evidence contrary to this interpretation; it instead appears that shortening the first retention interval, thereby decreasing the difficulty of the first retrieval, is indeed beneficial to long-term retention.

Complementing the identification of the main effects that determine a schedule's efficacy, this work also established a mediation effect wherein the previous retention interval was shown to be driving the ease of the next retrieval, which was in turn driving the ease of subsequent retrievals. This indicates that the effect of previous retention intervals partially acts through the ease of the following retrieval. Put another way, even though the eases of the

previous three retrievals were uniquely and positively predictive of the present ease, the retention interval preceding each of these retrievals partially determined their ease. These mediation effects are best visualized in Figures 15, 16, and 17. The influence of the previous retention interval on the following retrieval ease formed the theoretical basis of Experiment 2.

In Experiment 2 we demonstrated the successful manipulation of ease, whereby an adaptive scheduling algorithm kept retrievals closer to some target ease than predetermined, fixed schedules. This algorithm worked by modifying the expansion rate of a schedule in accordance with how far above or below the target ease a retrieval was. If a retrieval was easier than the target ease the expansion rate would increase, and if the retrieval was harder than the target ease the expansion rate would decrease.

This proof of concept opens the door for customized availability curves. Most prior works have tended to consider a learning schedule as simply a means to an end, where efficacy is measured by retrieval success on a single delayed retrieval attempt. This interpretation may be relevant in certain limited situations –such as a scheduled exam– where performance can reasonably be categorized as having a “learning period” and a “testing period”. The authors, however, argue that this is a somewhat simplistic interpretation of a learning schedule, and it makes more sense to define the ideal availability curve with respect to the specific goals of the learner. This interpretation does away with the concept of a learning and testing period. Instead, each retrieval is better described as part of an ever-growing learning period. This “testing” period, then, should simply be considered the next learning event schedule, even if accuracy on this retrieval attempt is more important than prior retrieval attempts.

To clarify the notion of availability being tailored to the goals of the learner, let us revisit the example of the student preparing for an exam and the language-learner attempting to achieve

fluency. For the student only concerned with exam performance and not retention after the exam, the ideal availability curve is the one that results in the highest availability peak at exam time, while minimizing the effort for the learner (i.e., reducing total schedule duration and/or retrieval attempts required to achieve this target availability). The language-learner, conversely, would likely desire an entirely different availability pattern, as they are interested in long-term retention and not just performance at some specific, predetermined time in the future. That is, you may have no idea when you will need to retrieve the Italian word for, say ‘eggplant’ (melanzana), but you would like to have a high probability of success whenever that situation arises. This language-learner, then, would likely prefer a schedule that achieves consistently high availability (rather than the one with the highest availability peak). The language-learner would also prefer a schedule that employs expanding retention intervals so as to create space to introduce new vocabulary words to their repertoire. Although Experiment 2 focused on a straightforward, constant target ease, this target ease could be modified to reflect the specific availability goals of the learner.

It is our hope that this work may prompt us to rethink some of the major assumption surrounding declarative learning (especially the notion of desirable difficulty), the factors that contribute to learning efficacy in both the short and long term, and the adaptive creation of schedules to support the unique availability aims of the learner.

Limitations

As with all work, there are a few limitations of note. Among these is the 14-point nature of the CASD scale, a key pillar of this work, formed through the unification of two 7-point Likert scales. As pictured in Figure 11, the CASD response values are not normally distributed, with skew towards either end of the scale. It is possible that the CASD scale is over-specified

and might explain a similar amount of variance (or more) with fewer response options. Based on the observed distribution, perhaps a 4-point scale would suffice, with two response values for correct retrievals and two for incorrect retrievals. A side benefit of this would be a reduction in cognitive load and response time for the learner.

Similarly, there may be concerns about the validity of the scale. While we tried to derive this CASD scale from the desirable difficulty literature and best approximate difficulty as a continuous function, we cannot claim that the CASD scale is the perfect method to do this. Alternatives certainly exist. For instance, a single bipolar scale could be imagined where the scale presented could be identical for correct and incorrect responses, rather than being dependent on retrieval accuracy. Other scales might be collected prior to the participant providing a response, potentially avoiding some of the “knew-it-all-along” effect. These options were considered but ultimately rejected for practicality concerns, as well as departing from specific claims made in the learning and difficulty literature. While we cannot claim the CASD scale to be perfect, we can state that it explains a great deal more variance in response likelihood than does response latency, the most widely used continuous, retrieval-level proxy for difficulty. In this sense it is a notable improvement on the current standard. Nonetheless, it is the author’s hope that future works might explore alternative, competing scales to estimate retrieval difficulty (and therefore the odds of retrieval success) as a continuous function.

Another limitation was the manipulation of stimulus difficulty, which was not significantly effective. In fact, the more complex stimuli were trending towards being better learned, rather than the less complex stimuli. This is perhaps due to the QR code format of the prompt. For the less complex stimuli, the prompts entailed 3x3 grids, which may have been difficult for learners to differentiate from one another. The potential for competition or

interference between these seemingly similar prompts may have contributed to the decreased –though non-significant– accuracy for these retrievals. The more complex 4x4 grids, on the other hand, may have created more memorable, identifiable patterns, paradoxically leading to increased accuracy. The manipulation of stimuli difficulty and potential interference among to-be-learned items merits further exploration.

Future Directions

Concerning other future directions for this work, we feel that it is important to explore the creation of custom, adaptive schedules in greater detail. For instance, it will be important to demonstrate the ability to target non-linear availability functions to meet the unique learning goals of different learners. Additionally, the efficacy of such schedules must be evaluated. The proof-of-concept described in Experiment 2 only targeted a specific, constant retrieval difficulty, and did not compare the adaptively generated schedules to one another. This type of comparison was deemed beyond the scope of the current work, and the experimental design was thus not created to afford this type of analysis.

Speaking more generally, maximizing the efficiency of declarative learning schedules is of vital importance to enormous numbers of people, ranging from students to language-learners to trainees (and many others). Anywhere formal learning is taking place, the findings of this work and others like it could be leveraged to reduce the time and effort required to learn declarative information to a satisfactory degree. And yet, relatively little effort is dedicated to ensuring that the staggering number of formal learners across the globe are not wasting their time studying inefficiently or in ways that are at odds with their specific learning goals. It is our hope that this work may provide some basis for decreasing this wasted or misguided effort though the

identification of concrete, readily applicable findings regarding the predictors of schedule efficacy and how to optimize declarative learning schedules.

REFERENCES

- Averell, L., & Heathcote, A. (2011). "The Form of the Forgetting Curve and the Fate of Memories." *Journal of Mathematical Psychology* 55(1): 25–35.
- Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger III, H. L. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology and aging*, 21(1), 19.
- Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General*, 127(1), 55.
- Bjork, R. A. (1994). Memory and metamemory considerations in the. *Metacognition: Knowing about knowing*, 185, 7-2.
- Bjork, E. L., & Bjork, R. A. (2011). Making things hard on yourself, but in a good way: Creating desirable difficulties to enhance learning. *Psychology and the real world: Essays illustrating fundamental contributions to society*, 2(59-68).
- Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. *Practical aspects of memory: Current research and issues*, 1, 396-401.
- Camp, C. J., Foss, J. W., O'Hanlon, A. M., & Stevens, A. B. (1996). Memory interventions for persons with dementia. *Applied Cognitive Psychology*, 10(3), 193-210.
- Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(6), 1563.
- Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 19(5), 619-636.
- Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological bulletin*, 132(3), 354.
- Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological science*, 19(11), 1095-1102.
- Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 14(3), 215-235.

- Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition*, 20(4), 374-380.
- Ebbinghaus, H 1885. *Über das Gedächtnis: Untersuchungen zur experimentellen Psychologie*. Duncker & Humblot.
- Gardiner, F. M., Craik, F.I.M., & Bleasdale, F.A. (1973). "Retrieval Difficulty and Subsequent Recall." *Memory & Cognition* 1(3): 213–16.
- Gates, A. I. (1922). *Recitation as a factor in memorizing* (No. 40). Science Press.
- Kang, S. H., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval?. *Psychonomic bulletin & review*, 21(6), 1544-1550.
- Karpicke, J. D., & Roediger III, H. L. (2007). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of experimental psychology: learning, memory, and cognition*, 33(4), 704.
- Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 37(5), 1250.
- Landauer, T. K., and Bjork., R. A. 1978. "Optimum Rehearsal Patterns and Name Learning." *Practical aspects of memory*: 52–60.
- Lawson, P.A, and Mayhorn, C.B. 2019. The Illusion of Desirable Difficulty and the Optimization of Declarative Learning Schedules. *NC Cognition Conference*. Raleigh, NC.
- Lee, J. L. (2009). Reconsolidation: maintaining memory relevance. *Trends in neurosciences*, 32(8), 413-420.
- Levenshtein, Vladimir I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*. **10** (8): 707–710.
- Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*, 15(3), 257-280.
- Nezlek, J. B. (2001). Multilevel random coefficient analyses of event-and interval-contingent data in social and personality psychology research. *Personality and social psychology bulletin*, 27(7), 771-785.
- Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words?. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31(1), 3.

- Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory?. *Journal of Memory and Language*, 60(4), 437-447.
- Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (Vol. 1).
- Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science*, 17(3), 249-255.
- Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences*, 15(1), 20-27.
- Spitzer, H. F. 1939. "Studies in Retention." *Journal of Educational Psychology* 30(9): 641-56.
- Stoet, G. (2010). PsyToolkit - A software package for programming psychological experiments using Linux. *Behavior Research Methods*, 42(4), 1096-1104.
- Stoet, G. (2017). PsyToolkit: A novel web-based method for running online questionnaires and reaction-time experiments. *Teaching of Psychology*, 44(1), 24-31.
- Wixted, J. T., & Ebbesen, E. B. (1991). On the form of forgetting. *Psychological science*, 2(6), 409-415.
- Wood, G. (1978). The knew-it-all-along effect. *Journal of Experimental Psychology: Human Perception and Performance*, 4(2), 345.