Title
Studies of Adaptive and Fixed Schedules in Factual and Perceptual Learning

Author
Mettler, Everett William

Publication Date
2013

Peer reviewed|Thesis/dissertation

University of California
Los Angeles

Studies of Adaptive and Fixed Schedules in Factual and Perceptual Learning

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Psychology

by

Everett William Mettler

2014

Abstract of the Dissertation

Studies of Adaptive and Fixed Schedules in Factual and Perceptual Learning

by

Everett William Mettler

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2014

Professor Philip J. Kellman, Chair

What conditions make learning efficient? Do adaptive methods – that use learner performance to arrange learning events – offer ways of improving learning? In this dissertation, I address these questions experimentally in human subjects across domains of factual and perceptual learning. Six experiments focus on methods of scheduling the order of presentation of items during a learning session with the goal of improving long-term retention. We introduce a method of determining spacing schedules using an adaptive, computer-based algorithm. Space between presentations is dynamically calculated as a function of a learner's response history and reaction time. This work is based on the spacing effect in memory: when learning a set of facts, scheduling time between repeated presentations or practice has been shown to improve learning. Larger gaps between presentations of an item contribute to increases in the strength of that item in memory. The primary experimental manipulation in these studies is a comparison between fixed schedules of practice and adaptive schedules of practice. Three crucial issues are tested. First, the effects of spacing in adaptive, computer-based schedules are compared to predetermined schedules that have fixed intervals of spacing.

Second, adaptive schedules are compared to schedules that are completely randomized, and the difference between schedules with and without dropout is compared. Finally, the benefits of adaptive scheduling are assessed using perceptual category learning. Results of the first set of experiments showed that an adaptive scheduling algorithm produced greater learning gains than fixed schedules, when the total number of presentations was limited. These gains were measurable after a one-week delay. Further, when fixed condition schedules were closely matched to adaptive scheduling in overall item spacing characteristics, adaptive schedules still outperformed them in terms of learning gains at immediate and delayed tests. Results of the second set of experiments, where learning was allowed to continue until learners met learning criteria, showed adaptive scheduling comparing favorably with random presentation schedules. Adaptive schedules where items were dropped after learning criteria were met produced greater learning efficiency (learning gains per trials invested), better than random presentation schedules with or without dropout These efficiency gains were maintained at a delayed test. In a final set of experiments, adaptive schedules increased learning efficiency during perceptual category learning reliably more than random schedules as well as schedules that were adaptive but included some initial blocking or massing of category exemplars. Adaptive scheduling also improved the fluency gains of students learning 3D chemical structure in an introductory community college chemistry course. The results in this thesis relate to state-of-the-art computer adaptive methods of teaching, as well as contemporary models of learning, memory, perception, categorization and human performance. These studies contribute to research in learning and memory, are of broad interest to educators who are concerned about student learning, and inform attempts to connect models of cognition with technology-based tools.

The dissertation of Everett William Mettler is approved.

James W. Stigler

Edward P. Stabler

Alan D. Castel

Robert A. Bjork

Philip J. Kellman, Committee Chair

University of California, Los Angeles

2014

To Oma

Appa

Squirrel

Duck

and Mooney .

Table of Contents

List of Figures

Acknowledgments

Chapter 4 includes work that has been submitted for publication: Mettler, E. & Kellman, P. (submitted), Vision Research. Conceived and designed the experiments: EM & PK. Performed the experiments: EM. Analyzed the data: EM. Wrote the paper: EM & PK.

The author gratefully acknowledges support from: the National Science Foundation (NSF) REESE Program award 1109228 and the National Institutes of Health (NIH) award 5RC1HD063338. Any opinions, findings, and conclusions or recommendations expressed in this work are those of the author and do not necessarily reflect the views of the the National Science Foundation, the National Institutes of Health or other agencies. Systems that use learner speed and accuracy to arrange learning events, as well as some aspects of the perceptual learning technology described herein, are covered by U.S. patent 7052277 and patents pending, assigned to Insight Learning Technology, Inc. For information, please contact either the author or Info@insightlearningtech.com.

Thanks to everyone who made this possible. Thanks especially to Phil for your patience and care, and for your unwavering dedication to good thinking and teaching. Also, your impeccable sense of timing didn't hurt. Thanks to my family for all your support. Thanks to everyone in the Human Perception Lab (especially Tim, Joel, Genna, KP and Rachel), everyone in CogFog, and all others, unnamed, who contributed with ideas, discussion or support. I couldn't have done it without you all.

Vita

2005	B.S. (Cognitive Science) UCLA, Los Angeles, California.
2006-2007	Research Intern, HRL Laboratories, Malibu CA
2009	M.S. (Cognitive Psychology), UCLA. Thesis advisors: Philip J. Kellman & Robert A. Bjork
2008-2012	Teaching Assistant, Department of Psychology, UCLA. Courses: Psych 85, 100B, 120A, 120B, 186C

Publications

Mettler, E. W., & Kellman, P. J. (2006). Unconscious discovery in concrete and abstract perceptual learning. Abstracts of the Psychonomic Society, 11, 17.

Mettler, E., Keane, B., & Kellman, P. (2008). Contour interpolation affects multiple object tracking. Journal of Vision, 8(6), 507.

Mettler, E., & Kellman, P. J. (2009). Adaptive sequencing in perceptual learning. Abstracts of the Psychonomic Society, 14, 88.

Mettler, E., & Kellman, P. (2009). Concrete and abstract perceptual learning without conscious awareness. Journal of Vision, 9(8), 871.

Keane, B. P., Mettler, E., & Kellman, P. J. (2009). Contour interpolation automatically directs attention in multiple object tracking. Journal of Vision, 9(8), 252.

Mettler, E., & Kellman, P. (2010). Adaptive sequencing in perceptual learning. Journal of Vision, 10(7), 1098.

Mettler, E., Keane, B., Erlikhman, G., Horowitz, T., & Kellman, P. (2011). Automatic feature-based grouping during multiple object tracking. Journal of Vision, 11(11), 287.

Mettler, E., Massey, C., & Kellman, P. J. (2011) Improving adaptive learning technology through the use of response times. In L. Carlson, C. Holscher, & T. Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 2532-2537). Boston, MA: Cognitive Science Society.

Thai, K., Mettler, E., & Kellman, P. J. (2011). Basic information processing effects from perceptual learning in complex, real-world domains. In L. Carlson, C. Holscher, & T Shipley (Eds.), Proceedings of the 33rd Annual Conference of the Cognitive Science Society (pp. 555-560). Boston, MA: Cognitive Science Society.

Keane, B. P., Mettler, E., Tsoi, V., & Kellman, P. J. (2011). Attentional signatures of perception: Multiple object tracking reveals the automaticity of contour interpolation. Journal of Experimental Psychology: Human Perception and Performance, 37(3), 685-698.

Mettler, E., Erlikhman, G., Keane, B., Horowitz, T., & Kellman, P. (2012). Further evidence for automatic, feature-based grouping in multiple object tracking. Journal of Vision, 12(9), 458.

Erlikhman, G., Keane, B. P., Mettler, E., Horowitz, T. S., & Kellman, P. J. (2013). Automatic feature-based grouping during multiple object tracking. Journal of Experimental Psychology: Human Perception and Performance.

0.1 Introduction

Purpose Scheduling spacing in learning has proven to be a highly successful, domain-independent strategy for increasing the efficiency and durability of learning. By 'spacing', researchers have typically meant the degree to which time or trials intervene between repeated presentations of an item. In short, relatively large gaps between presentations of an item contribute to increases in the strength of that item in memory. In contrast, excessive repetition of an item across small gaps appears to contribute much less to learning. There are also upper limits to the duration of spacing gaps, such that forgetting ensues if gaps are excessively long. The psychological and biological underpinnings of the spacing effect are still being researched, but in most empirical studies the effects of spacing have been found to be robust; effects are seen across numerous learning domains, in a variety of experimental and real-world contexts, across a wide range of temporal gaps, and across different learners.

Here we focus on methods of utilizing spacing effects in order to improve learning. The purpose of this project is threefold. First we hope to assess the benefit of computer-adaptive techniques that dynamically adjust spacing as a function of learners' ongoing knowledge acquisition. Second, the experiments are designed to provide a window into the general conditions and mechanisms that support successful learning and remembering. Third, we hope to apply the results of research to alternate learning domains using strategies that target the kinds of knowledge representations underlying real-world educational curricula.

CHAPTER 1

The Science of Learning and the Scheduling of Practice

1.1 Introduction

What does it mean to say that some method or procedure enhances learning? Why does the human mind benefit from some schedules of instruction or organization of material more so than others? When the mind benefits from strategies designed to increase learning, to what extent do those strategies directly target known mechanisms of mind? To what can we attribute gains in knowledge? These and other deep questions form the background for much research in learning and memory and for the specific research questions asked in this thesis. Here we isolate one effect from the extensive research literature in memory and learning — the spacing effect — and attempt to apply it using a novel, adaptive method of manipulating spacing during instruction. First we situate our research within related areas of investigation in the learning sciences. Next we review the spacing effect and discuss possible explanations of the advantages of spacing in memory. We then describe common proposals for utilizing spacing effects to enhance learning. We present an adaptive algorithm for scheduling presentations, one that uses ongoing indications of performance during learning to vary spacing, and we describe a series of experiments that investigate the differences between adaptive and fixed spacing schedules across a variety of conditions and learning materials.

1.1.1 Definitions of Learning

To talk about methods of improving learning, we must first ask, what is learning? In psychology, learning has been addressed in many ways. A summary by Hovland (1951) condenses half a century of attempts at definition. Hovland says: "It is desirable to include the trend of improvement in performance that comes about as a result of practice and to exclude changes in performance due to fatigue, sensory adaptation, and drugs, changes due to physical development (maturation) and artifacts in measurement, such as changes in rapport between subject and experimenter." Here 'performance' is defined broadly and the implication of 'practice' is perhaps too narrow: learning need not necessarily be deliberate nor the conditions of acquisition explicitly organized for learning to occur effectively (Reber, 1993). In humans, it seems almost no mental faculty or sensory mode is beyond the influence of this type of change. Learning can describe changes that may be affect single faculties of learning, such as how learning in perception changes the way information is extracted from the environment (i.e., perceptual learning - Gibson, 1969), and learning can be specific to a single sensory mode (e.g., visual learning). In addition to these definitions in psychology (and in addition to pedagogical and epistemological definitions not considered here) learning is also an abstract form of information processing as investigated in fields such as computer science and statistics (e.g., learning theory, see Valiant, 1984; Chomsky, 1975; or machine learning, see Duda & Hart 1973).

1.1.2 Goals of Learning and Metaphors of Mind

In order to explore the relative benefits of manipulations to the schedule of learning presentations, we adopt a set of modest learning goals. One goal of instruction is to maximize the quality and amount of information that a learner acquires. Another goal is to maximize the rate at which information is acquired. Finally, a third goal is to ensure that information is retained - maintained across useful periods of time. Thus, we conceptualize memory as a process by which information is attained, and where performance can be measured. Further

those improvements must be weighed against their long-term durability or retention. We recognize that these goals are only some of many potential goals, and that they require adopting the metaphor of 'mind as container', a metaphor that has often been used to criticize traditional instruction (Scardamalia & Bereiter, 1992; Lakoff & Johnson, 1980). Memorization of basic factual knowledge is sometimes considered a secondary goal to higher-order comprehension and conceptual understanding (Bransford, Brown & Cocking, 2000, Brown, Collins & Duguid, 1989). Despite these issues, learning of basic facts can still be an important part of instruction, and can be the starting point for the investigation of the learning of complex information. Investigations into more fundamental learning mechanisms are likely to provide the building blocks of a complete theory of human memory. Of course, debates over interpretations of learning will not simply disappear. Grappling with the diverse effects that higher-order learning and comprehension might have on rote memorization strategies is an important topic. Chapter 4 attempts to connect ideas about the scheduling of practice of facts to the scheduling of learning in complex material. Future researchers may explore even wider connections.

1.1.3 Human memory

Humans are able to store relatively complex pieces of information such as facts, images, experiences or self-generated thoughts - all or any of which we may refer to as memories. Further, we are able to retrieve those pieces of information from a vast database of personally held knowledge, often with seemingly little effort, and in little time. Though by no means perfect, human memory storage is very often precise, rapid, effortless, and long lasting. Modern research has taken pains to explain that memory does *not* resemble the processes of storage and retrieval familiar to electronic information processing devices like computers and recorders, there is no doubt that, in the main, human memory has many of the same characteristics that make mechanical analogues so remarkable. That humans are able to store and retrieve any information at all is of enormous consequence to the functioning of the species, as distinct from organisms that do not possess memory abilities. More vexing is

that we are far from understanding how the brain achieves its feats of information storage.

1.2 Models of Memory

To briefly review some theories of memory, there have been a range of approaches, from early and persistent attempts to describe the nature of learning as associative processes (Hebb, 1949; Thorndike, 1913; Hull, 1943), to researchers investigating the fundamental units of computation underlying complex mnemonic behavior (Landauer, 1975; Marr, 1970; Valiant, 2005), to modelers of memory attempting to delineate the law-like relations governing memory (Estes, 1950; Bush & Mosteller, 1955; Eich, 1982; Raaijmakers & Shiffrin, 1981). Theoretically and practically, the attraction of building a durable theory of memory has been strong - even researchers known for contributions in other fields have not been immune to the allure of investigating models of memory (Rock & Ceraso, 1964; Rumelhart, 1967; Simon, 1966; 1974). Memory has been central to the project of constructing unified theories of brain function (Anderson & Lebiere, 1998), and the practical implications have been large. Understanding memory processes would provide potential for relief for victims of brain and health maladies, as well as relief for the less serious travails of those with average memory, or even, the common student.

In this thesis we concentrate our efforts on discussing one heavily researched area, the 'spacing effect', in depth.

1.3 Spacing and Memory

1.3.1 The Spacing Effect

One longstanding finding in memory is that scheduling time or trials between re-presentations of an item increases memory for that item, relative to a schedule that has very short spaces or no space between trials. This effect, called the 'spacing effect', has been researched since early memory experiments by Ebbinghaus (1885/1913) and has been replicated and examined in

considerable depth by researchers since Ebbinghaus's time. There are actually three related effects of spacing, summarized by Crowder (1976): effects of repetition, distribution and lag. In this thesis we refer mainly to those spacing effects that describe the benefit of large 'lags' of an item across time, which we will continue to refer to as 'space' between presentations. To summarize again, the most useful spacing effect for our purposes is that which describes greater benefits to memory when the space between two practices of an item is increased.

The most basic example of the spacing effect in action is when a paired-associate (e.g., an english and foreign language word pair, henceforth called an 'item') is presented in a sequence of other items, and when the spacing between repeated presentations of each item is manipulated to include more or fewer of the other items. The number of presentations that intervene could be from 0, up to any number of intervening items. Findings in the spacing literature generally indicate less of an advantage to memory when the size of spacing delays is 0 or at a maximum, but better recall when spacing is intermediate (Glenberg, 1976).

1.3.2 Theoretical Explanations of Spacing

What can explain the gains in memory performance that result from spacing repetitions of an item across time? A few theories have been proposed to explain results, but in general, the specific mechanisms that underlie spacing effects remain elusive. Three theories are briefly described here, the last of which motivates much of this thesis.

Consolidation One potential explanation for the benefits of spacing in memory is that the time that elapses between presentations of an item causes greater consolidation of the memory traces that support retrieval of that item (Wickelgren, 1972). In essence, consolidation processes more deeply encode memories across time, and further presentations of an item inherit those more consolidated traces, benefiting later recall. Consolidation is consistent with cognitive and neuroscientific theories of memory; however, evidence for consolidation as a driver of spacing effects is less strong. One elegant study by Bjork and Allen (1970) showed that consolidation is not likely the best explanation for the benefits of spacing. Researchers

inserted either easy or difficult distracting tasks during a short or long interval between presentations of an item. Since difficult distracting tasks presumably disrupt consolidation, one would expect that, if spacing effects depend on consolidation, a difficult task would prevent any spacing effect. These researchers instead found the opposite: that difficult tasks during longer spacing intervals actually improved retention. Thus consolidation, while an important concept, does not appear to be the driver of spacing effects in memory. Instead, conditions of retrieval and encoding likely play a larger role.

Encoding variability Another explanation of spacing benefits is that circumstances of encoding are made more diverse when practices occur far apart in time. It is thought that diverse contexts at the time of encoding will be more likely to closely resemble the diversity of circumstances at final recall, thus facilitating recall performance. It is thought that the proximity of practice events affects the diversity of contexts during encoding, including such factors as: presence of other learning items, memory for previous presentations, and environmental or bodily status. By this account, learning advantages from spaced practice happen because encoding events naturally become more diverse as spacing increases (Raaijmakers, 2003; Kahana, 1996). These findings have even been incorporated in scheduling algorithms that attempt to optimize variability (and thus spacing) dynamically during learning (Mozer, Pashler, Cepeda, Lindsey, & Vul, 2009).

Unfortunately, a theory of encoding variability has difficulty accounting for some observed effects of spacing. So-called 'superadditivity' and 'non-monotonicity' effects are common to spacing but are not predicted by contextual variability models (Benjamin & Tullis, 2010). By superadditivity, researchers mean the general finding that repeated presentations increase learning even if the variability of encoding remains the same. Non-monotonicity describes the eventual decreases in the benefit of spacing that are seen as spacing gaps become large (Glenberg, 1976). Although encoding variability models might be able to account for these effects with some special exceptions or extra parameters, and while these phenomena do not exclude the possibility that encoding variability partially drives learning, it is not clear if

encoding variability exerts a sufficiently strong influence on memory mechanisms to explain spacing effects.

Retrieval difficulty A final set of theories considers the beneficial effect of spacing schedules on memory to be related to the amount of effort expended to retrieve items from memory. Importantly, more memory benefits are found when retrieval events are relatively more taxing — that is, more difficult in terms of the effort expended to correctly retrieve a memory trace. Less benefit is found when retrieval is easy. This 'retrieval effort' hypothesis has been demonstrated elegantly by a number of research programs and is compatible with comprehensive memory theories (Pyc & Rawson, 2009; Thios & D'Agostino, 1976; Bjork & Bjork, 1992). Results show that difficulty of retrieval can be induced in a number of ways, for example, by interleaving difficult tasks between retrieval attempts (Bjork & Allen, 1970), changing the amount of memory interference that retrieval attempts encounter (Storm, Bjork & Storm, 2010), or shifting a criterion number of retrieval attempts that an item receives before a test (Pyc & Rawson, 2009). Predominantly, however, retrieval effort can be induced naturally by stretching the intervals (duration of gaps) at which retrievals are attempted. This latter method leads directly to the technique of expanded retrieval practice which is discussed later.

To summarize briefly, difficult retrieval creates memory traces that are perhaps larger in number or elaboration, facilitating the eventual difficult retrieval of an item on a test. Retrieval difficulty is an important and widely supported theory of why spacing is beneficial for memory. Further, it is a hypothesis that is consistent with the predictions of the adaptive trial spacing algorithm that we introduce later and is a background assumption of the research in this thesis.

Physiological mechanisms Though we will not discuss them in depth here, it bears mentioning that recent investigations into the cellular basis of memory in neural systems point to specific biochemical mediators of memory formation, namely protein production and inhibition at both the presynaptic and postsynaptic level (Kandel, 2001; Jin, Kandel, & Hawkins,

2011). These findings have been extended to explore the possible neurochemical mediators and markers of the spacing effect. For instance, in Aplypsia, it has been hypothesized that protein kinase A (PKA) mediated long-term synaptic potentiation bears the hallmark of spacing effects in memory, and that patterns of serotonin induced activation consistent with timed spacing are superior at eliciting long-term sensitization than are massed patterns of activation (Naqib, Farah, Pack, & Sossin, 2011). Further, some research has begun to model and optimize the pattern of activation in these in vitro settings (Zhang et al., 2011; Philips, Kopec, & Carew, 2013). On the whole, these studies are still in their infancy; they are far from accurate models of how brain processes might store and recollect diverse, complex units of information, and they do not provide direct answers to ongoing research questions about the spacing effect. However the findings do provide some helpful guidelines. First, inasmuch as human memory processes partly involve similar neurochemical changes, these studies provide confirmation that mechanisms that drive the spacing effect are cellular and modular, operating largely beyond conscious intent or control. That is, spacing effects likely involve deep underlying brain processes that can operate in the absence of explicit control strategies. This relieves some of the explanatory burden from hypotheses that do not rely on explicit strategies, though it does not exclude the potential influence of strategy on those same underlying processes. Second, the potential for identifying parallels between models of spacing in synaptic potentiation and models of spacing in human learning are ripe. No matter whether higher-level behaviors are necessarily directly linked or constrained by lower-level processes, the possibility of comparing research at different levels is fertile and may spur useful productive results in each.

1.3.3 Other Constraints

It is important to note that the spacing effect we are focused on is not a universal one, and as indicated by Roediger (2008) cannot be described as a 'law' of learning[1]. Many laws of

[1]Within the domain of learning, there are a number of so-called laws such as the 'Law of practice' (Newell & Rosenbloom, 1981), 'Law of effect' (Thorndike, 1898), 'Josts Law' (Jost, 1897), and others that hold for a majority of cases. With a loose criterion the spacing effect could probably be considered a psychological

memory are known to break down when experimental variables like the type of material or variations in test or encoding conditions are present (Roediger, 2008; Jenkins 1979). The effects of spacing are no exception to this so-called 'relativity of memory' finding. What follows is a brief list of exceptional conditions that may warrant revisitation of the assumptions of spacing effects, either when designing research or building applied learning interventions that utilize spacing effects: Intentional vs. incidental learning; study vs. testing; feedback; types of learning materials or learning processes (e.g., perceptual, procedural, category learning); distinctions in memory processes (e.g., episodic vs. generic); and perhaps others. Future investigators who wish to apply the results of the current research to realistic educational settings should keep in mind areas where potential effects might break down, and if needed consult the appropriate literature.

The circumstances that can affect spacing are varied, and the effect does not always hold. Since spacing effects may vary with conditions, we have taken special care to ensure that the background conditions we have chosen (e.g., providing active recall trials rather than study trials) are most appropriate for revealing effects between the experimental variables of interest (namely, manipulations to the schedule of spacing). Thankfully, most of these choices also correspond to learning events that are known to improve learning (e.g., tests) and would be desired in any applied scheme for encouraging learning.

Ecological validity and diverse learning domains Little is known about how spacing effects might apply when the material to-be-learned consists, not of simple lists of words or facts, but of complex knowledge structures (for example, knowledge that is categorically or hierarchically organized, knowledge that is structured according to grammatical rules, or knowledge that requires the learning of temporally sensitive procedures or computations). Other questions arise if one particular to-be-learned item possesses features that are not completely independent from the total set of items being learned, or from the set of an entire learner's existing knowledge. These related problems address the issue of interactions

'law'; that is it meets the criterion of being an 'indispensable feature of the theoretical background' of the study of memory, and expresses 'the form of a regularity' and not its scope (Hanson, 1969).

between spacing effects and the specific mental representations that underlie various kinds of knowledge. Similar questions have been brought up by investigators looking to establish the applicability of spacing effects to realistic educational domains like mathematics (Rohrer, 2009)

Later in this thesis, we examine how adaptive schedules cope with information that is categorically organized, and learning material that targets learning processes other than fact memorization - specifically, perceptual category learning.

1.4 Schedules of Practice

1.4.1 Background

It has long been hypothesized that the reduction in the rate of forgetting across repeated learning opportunities should allow for the formulation of efficient schedules of practice; that is, schedules should maximize the amount of relearning at each practice and minimize the total amount of practice time (Tsai, 1927; Starch, 1927; Pimsleur, 1967; Landauer & Bjork, 1978). Combined with the ideas behind the spacing effect, a rich literature covers attempts to manipulate the duration of spacing, and the pattern of spacing practice, in order to maximize learning.

Research on fixed schedules can roughly be split into two paradigms: those studies that compare *classes* or *patterns* of spacing schedules (e.g., constant vs. expanding spacing), and those studies that deal with the parameterization of spacing schedules (e.g., the relative difference in recall when space is increased or decreased between presentations). In the first paradigm, a few basic classes of scheduling patterns have been analyzed for their effect on learning. For example, a classic finding is that expanding intervals — where delays between presentations of an item get increasingly longer — generate stronger memory traces than schedules where delays between items remain of a constant duration (Landauer & Bjork, 1978; Pimsleur, 1967; Storm, Bjork & Storm, 2010). In the second paradigm of spacing

studies, a small set of practice sessions is parametrically modified to assess the effect of duration on retention (Cepeda, Vul, Rohrer, Wixted & Pashler, 2008). Usually there are two practice sessions with a gap in-between, and a test of retention after a delay. The size of the gap and the delay to test are systematically manipulated to determine the relative influences of each duration on final retention. Unfortunately, in the latter type of research, where parameters of spacing are modified, the number of intervening gaps is few.

Combined, almost all current conclusions about spacing schedules in the literature are dependent on a small class of potential scheduling schemes. One promise is that adaptive schedules, such as that of Atkinson (1972) and Pavlik & Anderson (2008), and the adaptive system in the current thesis, may help to traverse more of the possible space of scheduling patterns than have been investigated with fixed schedules. In addition, adaptive schedules have rarely been directly compared to fixed schedules of practice — an important task that we return to later in this thesis.

1.4.2 Expanding Intervals of Retrieval

One highly advocated method of increasing memory retention across time is the method of successively expanding the intervals of spacing as practice accumulates during learning. This pattern of interval change during learning, known as 'expanding' retrieval practice, has been studied for nearly half a century (Pimsleur, 1967; Landauer & Bjork, 1978; Cull, Shaughnessy & Zechmeister, 1996).

Very recent research provides evidence for a substantial advantage of expanding the retrieval interval when material is highly susceptible to forgetting or when intervening material is processed between testing events (Storm, Bjork & Storm, 2010), conditions that apply to many formal learning situations.

Most explanations of the value of expanded retrieval intervals involve an underlying notion of learning strength. Learning strength can be thought of as a hypothetical construct related to probability of successful recall on a future test. When a new item is presented,

learning strength may be low, but it typically increases with additional learning trials[2]. The value of any new test trial varies with an item's learning strength. Specifically, evidence suggests that difficulty of successful retrieval is a crucial factor (Landauer & Bjork, 1978; Karpicke & Roediger, 2007; Pyc & Rawson, 2009). Pyc & Rawson (2009) labeled this idea the 'retrieval effort hypothesis': More difficult, but successful, retrievals are more beneficial. They studied the relation of number of successful retrievals to later memory performance, while manipulating the difficulty of those retrievals via number of intervening trials. Greater numbers of intervening trials led to better retention. One can summarize many of these findings by saying that the best time for a new presentation of an item is after the longest possible interval at which retrieval will still succeed. The idea is to stretch, but not snap, the retention interval.

Expanding retrieval as optimal for retention? A few studies have questioned the generalizability of expanding schedules of retrieval. In summary, it appears that when the material to be learned is somewhat easy, and when the second practice of an item is spaced enough to create difficult retrieval, future presentations of that item can be spaced by either constant or expanding intervals, with no difference to final retention (Karpicke & Roediger, 2007/2010). However, since the constraints that create those conditions may be very specific, it is unclear whether the benefits of expanding retrieval can be dismissed outright. Also, as pointed out by others (Pimsleur, 1967), expanding retrieval — even if equally as effective as constant spaced retrieval — is inherently more efficient (we return to the issue of efficiency in learning in Chapter 3). It also appears that when retrieval events are made difficult in general (by high levels of interference/competition with other items in memory, for example), expanding schedules result in greater learning than constant schedules (Storm, Bjork, & Storm, 2010) (We return to issues related to qualitative schedules of practice, e.g., the benefits of expanding vs. equal schedules in Chapter 2.)

[2]For simplicity we are describing a unitary concept of learning strength, though more elaborate notions of internal variables relating to memory strength have been proposed (e.g., Bjork & Bjork, 1992, propose two variables — memory strength and retrieval strength — to better explain many findings in the memory research literature.)

1.4.3 Open Questions

A paucity of researched schedules Although spacing effects have a long history of research, schedules of presentation have typically been narrowly investigated. That is, given the *space* of *potential* patterns of presentation across time, very few have been tested.

One limiting characteristic of research on patterns of spacing in learning is that most studies have chosen arbitrarily varying intervals of spacing - often primarily to satisfy methodological constraints in research - and have rarely given computational reasons why particular schedules are selected over others. As an example, research on patterns of spacing typically use approximately 3 or 4 spacing gaps between presentations of an item, commonly put into a notation that lists the number of intervening events (trials) between item presentations. For example, a '5 5 5' spacing schedule would indicate that 3 constant gaps, each of a 5 trial duration, would intervene between successive presentations of an item. Similarly, a '0 0 0' schedule would indicate no gaps between presentations (also known as 'massed' presentation). Crucially, some schedules correspond to so-called 'expanding' schedules of practice, where the gap durations are increased as presentations accumulate. For example, both a '1 4 10' and a '1 5 9' schedule would indicate expanding practice, with narrow gaps at the start of learning, and quite wide gaps near the end of learning. Existing schedules vary widely in the number and amount of delays, the size of delays, and the exact form of the pattern of repeated delays (e.g., 'expanding' or 'contracting'). Some examples of fixed schedules are printed here:

Parameterizing the study and retention intervals As mentioned above, although some studies have attempted to model the optimal size of spacing schedules, those studies have usually only attempted to do so with one spacing duration and one delay to retention test. Cepeda, Vul, Rohrer, Wixted, and Pashler (2008) analyzed the form of the function relating retention interval and spacing duration to recall at a retention test. They used

Type:	Equal	Expanding	Contracting	Description
	5 5 5	1 5 9	9 5 1	Mean delay = 5 (Landauer & Bjork, 1978; Karpicke & Roediger, 2007)
	2 2 2 3 3 3 4 4 4	1 2 3 1 3 5 1 3 8		Mean delays = 2, 3, 4 (Logan & Balota, 2008)
	7 7 7 7	0 3 7 18		Mean delay = 7 (Storm, Bjork, & Storm, 2010)
		1 2 3 3 5 8 5 8 13		Fibonacci subsets (Callender, 2010)
	1 5 9 5 10 15 15 30 45	5 5 5 10 10 10 30 30 30	9 5 1 15 10 5 45 30 15	Mean delays = 5, 10, 30 (Karpicke & Bauernschmidt, 2011)
	0 0 0			'Massing'

spacing intervals on the order of a few days to many weeks, and retention intervals on the order of weeks to months. Recall was worst for the shortest delays and best for slightly longer days, where delays that were maximal for retention increased if retention intervals were longer.

While the research in this thesis does not attempt to systematically adjust fixed patterns of spacing to find optimal patterns of retention, it is hoped that a direct comparison of fixed schedules with adaptive schedules might highlight the key differences between standard schedules and schedules that have been dynamically adjusted as a function of learner knowledge. Here we adopt standard fixed schedules that have been used in the literature in order to facilitate this comparison.

Key questions Might dynamic schedules that change parameters during a learning session compete better with the fixed schedules in the debates over qualitative schedules of practice? First we survey existing techniques to schedule practice during learning, both interactive, fixed, and adaptive. Then we introduce an adaptive algorithm that determines optimal spacing delays during learning.

1.5 Learning Technology and Other Applied Approaches to Scheduling

While the history of techniques designed to improve learning extends beyond just those that manipulate the order of practice, some forms of augmented instruction have been specifically designed to alter the pattern or schedule of practice in order to engender durable, long-term learning. Since one primary purpose of examining the benefits of adaptive schedules of practice is to develop learning technology that can be employed in classroom settings, we review a few examples of prior and existing learning technology, examine their theoretical motivations and relate them to issues of scheduling in memory.

1.5.1 Non-Computerized approaches

Spacing effects have already been leveraged in a number of existing curricula and learning systems. Here we describe non-adaptive implementations of the spacing effect, which can be divided into either fixed spacing, or expanding spacing types.

Fixed spacing Interestingly, spacing effects are naturally implemented in a number of learning strategies such as flashcards, have been purposefully integrated into static curricula (e.g., the Saxon mathematics curriculum), and have been documented in foreign math textbooks that encourage distributed practice of previously learned math concepts (Stigler, Fuson, Ham, & Kim, 1986).

Expanding spacing Some existing methods of scheduling presentations during learning already utilize the hypothesized benefits of expanding retrieval practice. One famous method is the Leitner method of flashcard use (Leitner, 1972, Mondria & Mondria-De Vries, 1994). In this system, flashcards are separated into groups, and then learned to some criterion of successful retrieval. When groups have been successfully learned, they are placed in a re-practice queue, ordered by the time since learning. The system ensures that previously learned stacks of cards are recycled (re-practiced) at longer and longer delays since original learning. This method naturally implements a version of expanding practice, although it is easy to see that the durations between representations are arbitrarily wedded to such constraints as stack size and total number of cards.

Pimsleur (1967) also described a method of expanding the delays between re-presentations of an item — a method he called 'graduated interval recall'. Pimsleur tuned his method to work optimally for the learning of foreign language vocabulary and conversation, and his system elegantly scales upward from the learning of individual words and short phrases to the development of complex conversational skills. Although there is no published work demonstrating exactly how he determined the length of the delays between presentations of items, his published thoughts about the method are among the earliest describing the benefits of an expanding schedule of practice. Anecdotally, his system is effective at training basic foreign language ability in a short amount of time.

Both of these expanding methods of spacing are still in use today, and the Pimsleur method has been commercially successful, though neither system has been well researched. It remains to be seen whether these methods would fare as well as alternate expanding, adaptive or fixed learning schedules in experimental comparisons.

Learner-controlled schedules Some work has looked at learner-controlled schedules of instruction (Woodson, 1974; Ciccone & Brelsford, 1976). Major findings tend to discourage the use of learners' own judgments in scheduling decisions (e.g., flashcard learning, Kornell & Bjork, 2007), and metacognitive awareness of learning has generally been deemed poor

(Kornell & Bjork, 2008; Kornell, Castel, Eich & Bjork, 2010).

1.5.2 Computerized approaches

Computer assisted instruction (CAI) As stated before, research has shown that spacing can affect changes in learning strength of items as learning progresses. Some previous adaptive approaches have relied on accuracy and trial history to predict learning strength, either in a Markov model estimating transition probabilities between different states of retention (e.g., Atkinson, 1972) or more elaborate models of learning (Pavlik & Anderson, 2008; Wozniak & Gorzelanczyk, 1994).

Atkinson (1972) designed a scheduling algorithm that used a Markov model to track learning of vocabulary items during interactive learning sessions. A Markov model is a mathematical description of state transition probabilities for a simplified system where the distributions of probabilities of future states depend only on the current state and not on past states. The model was used to track the probability that a learning item was in hypothetical states that reflect different degrees of learning (e.g., unlearned vs. well learned). Atkinson conducted an experiment to compare the following scheduling strategies: self selection of items, random selection of items, selection of items using a Markov model with individual parameters for each item, selection of items using a Markov model with parameters estimated for all items.

Atkinson had participants learn German-English word pairs. Atkinson's results showed that learners had greater retention when they were trained with an algorithm that had prior parameters for each item; less performance when prior parameters were determined for the entire set; and somewhat less performance for items that had been randomly presented or that had been self selected by learners. Results suggested that using an adaptive model – in addition to determining ideal item parameters using a prior learning session – was highly effective in improving memory performance for realistic learning material.

Pavlik & Anderson (2008), also demonstrated a model that utilizes spacing effects to

boost learning. They reported strong learning results – better than with Atkinson's (1972) approach, using a detailed cognitive model of acquisition, based partially on ACT-R (Anderson & Lebiere, 1998). Pavlik and Anderson's model relied on using prior studies to acquire learning parameters for individual items and comparable learners, before an adaptive session was run.

Motivating an adaptive approach that relies on response speed and calculates optimal spacing delays What none of these prior approaches provide is an ongoing model of learning strength that can be related to the degree of potential learning benefit conveyed by spacing delays. Crucially, the benefit of spacing delays can be computed as a function of learning strength. If learning strength is low, spacing delays must be short in order to support retention; if learning strength is high, spacing delays will need to be long owing to shallower rates of forgetting. One important way to measure learning strength during learning is to measure response speed. Response speed provides evidence for the degree of current learning strength; if responses are fast, learning strength is high, and if responses are slow learning strength can be assumed to be low[3] (Benjamin & Bjork, 1996; Pyc & Rawson, 2009). Using response speed, and leveraging what is known about spacing effects, we aimed to construct an algorithm that could automatically generate spacing delays for learners as learning progresses.

In this thesis, we assess this algorithm in a variety of ways. To determine the power of adaptive spacing at the level of individual delays and presentations, we compare adaptive scheduling to fixed scheduling. Expanding the scope of that investigation, we assess the power of adaptive schedules that incorporate learning criteria in longer durations of practice. Finally we assess the effect of the algorithm in different learning domains. Namely, we test adaptive scheduling in the fact learning domain as well as in *perceptual learning* - the pickup of structural information in perceptual tasks.

[3]With some exceptions owing to the contrasting benefits of recent practice on learning strength - with recent practice, reaction time may remain fast even though learning strength is low (see Bjork & Bjork, 1992 for a more detailed theory of learning strength).

1.6 Adaptive Response-Time-Based Sequencing (ARTS)

Here we present a scheme for dynamically adjusting the schedule of practice: adaptive response-time-based sequencing (ARTS). Importantly, this scheme uses both the accuracy of a learner's responses as well as their response *speeds* to calculate new spacing gaps between presentations. Basing adaptive schemes on both accuracy and response-times offers a more direct way to assess learning strength for individual learners and items in an ongoing manner. In our system, retention intervals expand as an inverse function of response-time (for accurate responses), such that faster responses automatically produce longer recurrence intervals. Consistent with many studies and models, the approach assumes that learning strength is reflected in response-times (Benjamin & Bjork, 1996; Karpicke & Roediger, 2007; Pyc & Rawson, 2009).

1.6.1 Model Detail

ARTS applies principles of learning to all learning items simultaneously in a *priority score* system, in which all items are assigned scores indicating the relative importance of that item appearing on the next learning trial. Priority scores for each item are updated after every trial, as a function of learner accuracy and RTs (reaction times), trials elapsed, and in view of predetermined mastery criteria. Learning strength is assessed continuously and in some implementations, cumulatively, from performance data. The most straightforward version of our sequencing algorithm chooses the highest priority item for presentation on each learning trial. Adjustable parameters allow flexible and concurrent implementation of several principles of learning and memory. One important principle is that the retention interval (spacing gap) automatically increases for an item as its learning strength grows.

Priority scores are dynamically updated after each trial. In many applications, initial priority scores are given to all items, and an item's score does not change until after it is first selected for presentation. This establishes a baseline priority for feeding in new items that may be balanced against changing priorities for items already introduced. Parameters

may be set to favor recurrence of new items, items already seen, or combinations of the two.

The sequencing algorithm is flexible; it may utilize any equations relating elapsed time or trials, accuracy, and RT to the priority for presentation of an item on a given learning trial. When any particular function of these variables is used, parameters may be adjusted to suit particular learning contexts and even individual learners. We describe here a characteristic priority score equation that allows implementation of several key principles of learning and has proven highly effective in our prior research. The Priority Score for item i (P_i) is given by:

$$P_i = a(N_i - D)[b(1 - \alpha_i)Log(RT_i/r) + \alpha_i W] \tag{1.1}$$

where:

$$\begin{aligned} N_i &= \text{number of trials since item i was presented} \\ D &= \text{enforced delay constant (trials)} \\ a, b, r &= \text{weighting constants} \\ \alpha_i &= 0, \text{ if learning item was last answered correctly} \\ &= 1, \text{ if learning item was last answered incorrectly} \\ W &= \text{priority increment for an error} \\ RT_i &= \text{response-time on most recent presentation of item i} \end{aligned}$$

Parameters In addition to the variables in the sequencing equation, there is a parameter P_d that describes the default priority score for items that have not been introduced. Typical parameters settings for the algorithm are listed in the Appendix in Table A.1. Generally, there are few changes needed to the parameters to accommodate differing learning material and learners, however D (delay constant) and r (RT weight) are commonly modified to adjust the global size of spacing delays and the gap between presentations after incorrect responses.

1.6.2 Other Characteristics

Rapid Reappearance of Missed Items The system ensures rapid re-presentation of items answered incorrectly by the assignment of a high priority weighting increment (W). The binary variable α_i is used to activate one or the other part of the equation, depending on whether the last trial response was correct or not. If correct, α_i is set to 0, and priority becomes a function of RT. If incorrect, α_i is set to 1, and priority increment W applies to the item. With ordinary parameter settings, the error increment W will exceed all initial priority score assignments, as well as the highest priority that may result from a slow, correct answer. However, reappearance of missed items is still subject to enforced delay (see below). With typical parameter settings, a missed item will tend to have highest priority, once it passes the enforced delay.

Interleaving / Enforced delay To prevent presentation of an item while its answer remains in working memory (Karpicke & Roediger, 2007; Taylor & Rohrer, 2010), the system is configured to prevent the presentation of the same item on consecutive trials. The parameter N_i and constant D determine the enforced delay, because (N_iD) is a global multiplier in the equation. A value of 2 is typical for D, and N_i represents number of trials since last presentation of item i. Thus, the overall priority of item i will be negative on the trial immediately following the error (because $(N_iD) = -1$). On the next trial, the priority will be 0 (because $(N_iD) = 0$). For both negative and zero values, the priority for re-presentation of item i will be lower than all learning items having positive priority values. From then on, the priority for a missed item ensures imminent presentation, as its priority increment W grows proportionally to the number of elapsed trials since last presentation.

Dynamic spacing based on RT The system can use various functions of RT but typically produces large priority scores for slower rather than faster accurate responses. In the priority equation: For an item answered correctly, $\alpha_i = 0$, and the part of the equation involving RT is activated. RTs for inaccurately answered items are not considered meaningful. For

correctly answered items, a log function of RT is used. In this arrangement, longer spaces between presentations of an item arise automatically as the learner gives faster (accurate) responses.

Retirement criteria In some of the studies proposed here, we use 'retirement', or removal of an item from the learning set, based on attainment of learning criteria. Pyc & Rawson (2007) called this 'dropout' and found evidence that greater learning efficiency can be achieved with this feature, especially in highly demanding learning situations. In addition, because Kornell & Bjork (2007) have shown that learners are quite poor at deciding which items to remove from a learning set, allowing the learning system to control dropout is a reasonable feature. The learner has to answer an item correctly and under a criterion response-time on consecutive (widely spaced) presentations to retire that item. Requiring several, consecutive, fast responses to an item automatically ensures stretching of retention intervals. Thus, a retired item will have been answered quickly and accurately several times across long delays before being retired. (In the experiments where total number of presentations is fixed (Chapter 2: Exps. 1 & 2), the retirement feature is not used.)

Our approach concurrently incorporates a number of learning principles supported by recent research. The ARTS system is built around short interactive trials, an approach supported by considerable evidence indicating that interactive 'testing' trials, in which the learner makes a response, are highly effective in learning, more so than passive presentations or 'study' trials (Carpenter, Pashler, Wixted & Vul, 2008; Karpicke & Blunt, 2011). The use of systematic mastery criteria, including speed, assures both comprehensiveness and fluency in learning. As cognitive load is an important limiting factor in learning (Chandler & Sweller, 1991), it is important that items that are foundations for later learning be mastered to a reasonable degree of fluency. Finally, the rich stream of performance data accumulated by the ARTS system enables continuous assessment by instructors, and also provides several forms of learner-directed feedback, which can support specific increments in learning and sustain motivation.

1.6.3 Comparison with Past Systems

Empirical comparison with Atkinson Markov Model (1972) We compared ARTS to Atkinson's (1972) system. To recall, Atkinson's system was based on a Markov model tracking strength of learning items. Presentations were chosen as a function of probabilities of transitioning between three hypothetical learning states – unlearned, temporarily learned or permanently learned. The algorithm attempted to select items that would have the highest probability of moving from an unlearned or temporarily learned state into the permanently learned state if tested and studied on the next trial. Previous learning data were analyzed to determine the model's initial parameters, including learning and forgetting rates and prior knowledge. Atkinson successfully used his model to improve learning of German-English vocabulary pairs (and used related systems in a variety of domains; for a review, see Atkinson, 1968). Performance, as measured by recall on a delayed post-test, was superior to random presentation.

We compared the ARTS system with a version of the Atkinson model using material that consisted of names and locations of countries on a map of Africa (Mettler, Massey & Kellman, 2011). To implement the Atkinson condition, item parameters were estimated using data from a previous experiment, in a manner similar to that in Atkinson (1972). No prior information was required for implementation of the ARTS system.

Method Participants were randomly assigned to two learning conditions. One group received training using ARTS. The other group received training using the Atkinson scheduling algorithm. Each group of subjects took a pre-test in which they were asked to identify 24 countries on a map of Africa. We used countries whose location was relatively unfamiliar (e.g., Djibouti, but not Egypt). On each trial, a country was highlighted on the map, and participants were asked to choose its name from a list of 24 country names. Countries were presented individually and no feedback was given.

Results The results showed that adaptive sequencing based on response-times and accuracy can produce substantial enhancements in learning relative to other methods. The ARTS system was 54 percent more efficient on immediate post-test based on trials and 76 percent more efficient based on time than the Atkinson (1972) approach, and these differences were equally evident on delayed post-test (see Figure 1.1). The Atkinson condition tested in this

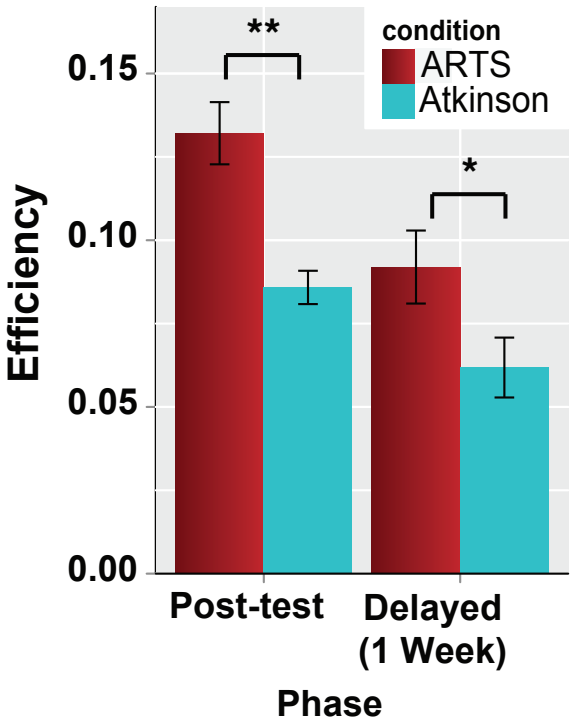

Figure 1.1: Efficiency for ARTS and Atkinson scheduling algorithms at immediate and delayed (1-week) post-tests. Efficiency equals improvement in number of post-test items answered correctly per trial of training. Reprinted from Mettler, Massey and Kellman (2011).

study has been shown in prior work to offer substantial improvement over random schedules of presentation (Atkinson, 1972), so we might infer that the ARTS system would outperform random schedules. However, we go further in this thesis, to demonstrate the power of adaptive schedules directly compared with random schedules of practice (Chapter 3).

1.6.4 Motivating Issues

What explains benefits of adaptive spacing? What are reasons that adaptive schedules could confer benefits during learning? Three factors come to mind. First, adaptive schedules might benefit learning because they take into account random and unexpected variation in learning. This variation could manifest in different facets of the learning process. For example the typical accumulation of knowledge, or change in learning strength on a given trial, may suffer from chance fluctuation. Alternatively, there may simply be random variability in performance rather than in the knowledge of an item. An adaptive system may account for these types of errors better than a system that has pre-determined and fixed amounts of spacing between trial presentations. Secondly, a learning scheme that separately tracks a learner's performance on individual items may perform better than one that does not. Such a scheme could track individual items in addition to individual learners. Third, a learning scheme that adjusts to the dynamics of the set of items to be learned — as an example, changes spacing as a function of the number of items that are in a working set of currently practiced items — may better cope with emergent properties of the learning session.

1.6.5 What is the Range of Benefits that Adaptive Spacing can Confer?

In what conditions will an adaptive algorithm help learning? With what stimuli and what modes of processing? Are there limitations to the complexity of material that can be learned? Are there instances where adaptive approaches would not benefit learning? Questions about the range of materials and conditions that are affected by adaptive learning will be addressed.

In this thesis we test hypotheses about the usefulness of adaptive schedules over that of fixed, spaced practice, through a variety of experiments. First, we compare adaptive scheduling with fixed scheduling (Chapter 2, experiment 1). Next we attempt to assess the locus of learning benefits in adaptive learning (Chapter 2, experiment 2). We also compare adaptive schedules under conditions that are more prolonged, matching conditions

of study and outcomes of learning that would be expected in real-world settings (Chapter 3, experiments 1 & 2). Finally, we assess whether the benefits of adaptive scheduling also apply to perceptual learning – an alternate learning process that may be influenced by spacing techniques (Chapter 4). Throughout we assess the theoretical reasons that adaptive learning techniques might bring benefit to memory and learning processes.

CHAPTER 2

Adaptive Schedules vs. Fixed Spacing Schedules

2.1 Introduction

Background This set of experiments compares adaptive schedules to fixed schedules of practice. The motivation for these studies is twofold: First, the literature that addresses patterns of spacing in memory typically considers a set of schedules that differ in the prescribed duration of delays separating presentations of an item. Fewer studies have examined schedules that adapt, or change duration during practice, and no studies have directly compared schedules of fixed delays to schedules that algorithmically adapt delays during the course of learning. Second, studies that examine the character of learning schedules that change delay duration over the course of multiple presentations typically compare categorically different schedules, such as fixed expanding vs. fixed equal presentation schedules, joined by a common measure such as mean delay size, rather than attempt to match schedules to actual needs of the learning material. The current set of studies systematically assess an adaptive schedule that administers multiple, adaptive delays for each item during learning. In doing so we aim to answer a set of related questions. First and generally, do adaptive schedules of practice outperform fixed schedules (e.g., of either the constant delay or expanding delay type)? We can test this by directly comparing fixed and adaptive schedules. Second, if adaptive schedules are better, what could be a specific reason that adaptive schedules confer benefit over fixed schedules? For instance is adaptation to individual learners more powerful than adaptation to individual items? We can test this by implementing a set of fixed schedules of practice that are either more or less congruent with the required spacing dynamics of individual items or individual learners.

We might expect that an adaptive system which can accurately measure a learner's current knowledge will be better prepared to react to variability in the learning process, variability among learners, and variability among items, than any fixed schedule. We might also expect adaptive schedules to be better than fixed schedules that have been tuned to individual differences among items. Again, as described previously, we use an adaptive algorithm that computes spacing intervals as a function of response time and accuracy. Since response time on correct trials is thought to be a direct indication of an item's learning strength in memory, it is hypothesized that schedules of practice which use response time and adapt dynamically to a learner's current learning strength will perform better than any schedule that does not respond to learning strength.

Prior work In terms of a comparison of adaptive algorithms and fixed schedules, the present work is the first known attempt to compare the two types of schedule. A growing literature explores methods of scheduling based on adaptive techniques[1], and a separate large literature addresses issues related to scheduling of a few fixed trials of practice; however no study at present has attempted to compare both fixed and adaptive schedules, and to assess the relative benefits of spacing to long-term learning in each.

2.1.0.1 Debates about scheduling

A few studies focusing on fixed schedules bear directly on the work attempted here. First, some studies have attempted to show the benefits of particular qualitative schedules of practice, as mentioned in Chapter 1 (Karpicke & Roediger, 2007; Landauer & Bjork, 1978; Storm, Bjork, & Storm; Carpenter & Delosh, 2005). Second, some studies have attempted to address the range of the power of spacing intervals over repeated presentations (Karpicke &

[1]Techniques do not always agree on the types of learning goals to achieve. For instance some techniques aim primarily to reduce the amount of total time spent practicing items, thus targeting the learning of items most likely to benefit from extra practice in the long term, but at a sacrifice to items deemed too difficult to learn quickly (Pavlik & Anderson, 2008). Few adaptive schedules attempt explicitly to maximize the duration of spacing delays to optimize learning for each item, and we know of no other techniques that rely on ongoing measures of response speed during learning.

Bauernschmidt, Pashler, Zarow, & Triplett, 2003), as well as interactions with features of learning such as dropout and learning criteria (Pyc & Rawson, 2009). Though there is persistent controversy throughout this body of work, there has been much progress in determining the range of benefits that come from distributed scheduling, as well as some old and new questions.

Is expanding practice optimal for retention? One issue that has arisen is a debate over the relative benefits of expanding vs. equal intervals of practice. Despite intuition and some evidence that expanding retrieval practice is the most effective distributed scheduling technique (Landauer & Bjork, 1978; Pimsleur, 1967), other evidence indicates that, to the contrary, there is no difference between expanding and equal schedules of practice (Carpenter & DeLosh, 2005; Karpicke & Baeurnschmidt, 2011) or even that equal interval practice is superior to expanding practice (Karpicke & Roediger, 2007; Logan & Balota, 2008) or superior at a delay (Cull, 2000). For example, Karpicke and Roediger (2007) showed that expanding interval practice (1-5-9 intervening trials between 4 presentations) led to poorer performance vs. equal interval practice (5-5-5 intervening trials) at a delayed test, whether or not training had feedback (without feedback recall accuracies were 0.33 vs. 0.45 after a two day delay, and with feedback recall accuracies were 0.49 vs. 0.60). These findings are in contrast to studies that have shown effects in favor of expanding practice (Landauer & Bjork, 1978; Storm, Bjork, & Storm, 2010).

The lack of demonstration of the superiority of expanding schedules, as well as the possibility of their inferiority, has generated a few explanations. One rationale is that, in any given schedule of multiple, repeated practices, the benefits to learning come primarily from the initial intervals of practice. In some cases, the smaller initial intervals of expanding practice help learning - such as when learning strength is low for poorer learners or for difficult material (Cull, Shaughnessy, Zechmeister, 1996). In other cases, however, smaller initial intervals are insufficiently short, and do not add appreciably to learning. In these cases, an equal practice schedule whose initial intervals are long (in fact uniformly long through-

out), do in fact benefit learning. Without wasted opportunities on substandard delays, equal schedules benefit from a greater number of longer delay intervals. A further possibility is that, after a few initial intervals, later intervals in these schemes have very little effect on learning (Karpicke & Roediger, 2010). A second, somewhat competing rationale is that the power of different schedules interacts with the difficulty of intervening learning material (Storm, Bjork & Storm, 2010). With more difficult material, the power of expanding practice is clearly shown.

We add here a third possible rationale that we think can help explain and unify the disparate effects and embattled theories in these prior studies. This rationale is simply that, since expanding schedules are not, on their own, powerful enough to predict when and where learning strength might increase or decrease during a learning session, there will necessarily be both instances in which expanding schedules are better and instances in which expanding schedules are worse than equal schedules. It should be remembered that the hypothesis sanctioning larger and larger delays between practice hinges on both the difficulty and the success of repeated retrievals. There are thus two maladaptive events that could contravene the possible benefits expanding delays bring to learning: retrievals that fail due to exceedingly long delays, and those easy retrievals that fail to add much to learning strength. Equal intervals do not have the former drawback, and are usually designed not to have the latter because delay schedules are usually chosen so that average delay intervals are not beyond the limits of reasonable practice. A casual analysis of results in Karpicke and Roediger (2007) seems to confirm that schedules could not maintain difficulty of retrieval across the expanding schedule, as indicated by relatively higher response speed across each presentation in the expanding than the equal condition. Response speeds declined proportionally more quickly across the final interval of the expanding condition than the equal condition, suggesting that retrievals could have been made more difficult in an expanding scheme by expanding delays toward the end of learning - but they were not made so. All of this is to claim that, in order for expanding schedules to benefit learning they need to possess delays that closely track ongoing learning strength, a feat made difficult when learners and learning material

vary. To reiterate, it is very likely that expanding schedules are powerful drivers of learning - and especially efficient learning - but that when fixed (pre-determined), expanding schedules alone cannot always make up for differences in learners' ongoing learning strength or for differences across learning material. Contrary to Karpicke and Roediger's (2007) comment, 'Considering the widespread belief in the utility of expanding retrieval, it is surprising that there is not a larger base of research with consistent evidence showing expanding retrieval practice to be the superior spaced practice technique for improving long-term retention,' we claim instead that the lack of demonstration is unsurprising given the lack of flexibility of the match between the exact parameters of the schedule and the variety of materials and conditions that experimenters have utilized in testing [2]. It is possible that, in theory, that claims about expanding schedules remain sound. One purpose of the experiments in this chapter is to see if any evidence for the benefits of expanding practice might lie in the pattern of schedules that are generated in adaptive schedules of practice.

What learning events are optimal? Aside from studies that directly compare two qualitative types of schedule, some studies have attempted to ask questions about the form of optimal spacing delays regardless of the qualitative schedule, and about the character of the learning events within schedules. That is, these studies attempt to predict the relative size of spacing advantages based on spacing patterns and trial features. For instance, some have asked what the general size of optimal spacing delays should be (Karpicke & Bauernschmidt, 2011), or how spaced practice differs depending upon whether practice is successful or unsuccessful during learning (Pashler, Zarow, & Triplett, 2003). Karpicke and Bauernschmidt's work suggests that optimal spacing schedules are those that possess the largest absolute delay size. In a comparison of schedules with average delay sizes of either 5, 15 or 35 trials, the best recall performance was found for the spacing schedules with delays that averaged 35 trials, regardless of the qualitative pattern of those schedules [3]. Items were presented

[2] As Cull et al. indicate, 'boundary conditions' do much to mediate the expanding retrieval spacing effect.

[3] Karpicke and Baeurnschmidt (2011) also compared equal, expanding, and contracting schedules but found no difference in recall for them. Again, to prove that expanding intervals were consistently engaging difficult recall they looked to response time as an indicator of retrieval difficulty and showed that the slope

with feedback, and there was relatively little forgetting at each delay duration, except with the largest (35 trial average) schedules where recall after each delay interval was consistently about 90%.

Related to this work, Pashler, Zarow and Triplett (2003) showed that the effect of the size of delays between two presentations moderated the effect of recall at a delayed test, regardless of the recall performance at the second presentation. In their study they gave participants paired associates (foreign vocabulary items) at two presentations separated by a parameterized delay interval of between roughly 1 and 32 trials. At each presentation participants were given feedback showing the correct answer. Interestingly, performance at the second presentation was highest for the shortest delay size (around 90% correct) and lowest for the largest delay size (around 50% correct). More interestingly, delayed recall performance (the next day) was highest for the largest delay size (still at 50% correct) and lowest for the shortest delay size (well below 50%). Since there did not seem to be any indication of the classic 'non-monotonic curve', found in most spacing effect research, a second experiment also extended the range of possible delays to 96 trials. Even with 96 trial delays and alternate learning material, it still appeared that, consistent with Karpicke and Bauernschmidt (2011), that the largest delay sizes were most beneficial to learning. Their conclusion was that errors – combined with feedback after errors – may actually be facilitating rather than preventing spacing effects. [4]

It should be obvious that both of these studies' findings are contrary to the theoretical framework we have been espousing, namely that in schedules of repeated practice it is important to have difficult but successful retrievals. For the Karpicke and Bauernschmidt work, there are two issues. First is that their method required items to be learned to criterion *before* spacing schedules were tested. Second, as in other prior work, we feel that the patterns

of response time decrease was shallowest for expanding retrieval. However a casual look at their results for response time reveals that their slope estimates were too biased in favor of the response times at the final presentation, where responses times were in fact highest for the expanding condition; for the majority of the prior presentations, equal presentation schedules had higher response times.

[4]Some criticisms immediately come to mind: For instance the construction of schedules with some highly short and some extremely long delays is likely to bias items with long delays into positions of early or later learning - thus instigating corresponding primacy and recency effects.

of expanding practice did not map to item characteristics and learner characteristics in an optimal way - and that the benefit of absolute delay sizes will eventually have its limit (this is hinted at in the analysis of Pashler, Zarow, and Triplett, 2003). For the Pashler, Zarow, and Triplett work, it could be that in schedules where there are only two presentations of an item, failure after long delays encourages encoding processes or strategies that try to 'make-up' for the lack of future presentations. In addition, there could be an interaction with the kinds of spacing delay optimal for certain material, such that errors are harmful if spacing delays are short, but beneficial if spacing delays are excessively long. Clearly there were differences in the method of these two studies and the work we are presenting here. We attempt direct tests of comparing adaptive schedules to fixed schedules, across 4 presentations and without prior learning to criterion. Some of our analysis will go to show that successful retrievals after each practice are in general important to generate better learning, and that conversely, 'snapped' or failed intervals are not as beneficial for learning.

Key questions and motivation In a comparison of adaptive schedules to typical fixed schedules of practice, we expected to find greater learning in the adaptive condition than in either of the fixed conditions, but we also hoped to weigh in on debates about scheduling in the literature such as the merits of expanding vs. equal delay fixed schedules. Further, we hoped to answer a small set of questions: First, we wished to know the gross capabilities of the adaptive algorithm performing in the absence of learning criteria across a fixed set of presentations. Second, we wished to observe differences between characteristics of the learning session between fixed and adaptive items. Finally we hoped to examine the resulting adaptive trial data, to reveal hitherto unknown aspects of learning schedules that might inform future fixed or adaptive schedule designs. Of course, in the main it was expected that an adaptive algorithm would outperform fixed schedules of practice, but we hoped to reveal more precise reasons for this advantage through experimentation.

It is worth asking what we would learn if we instead found deficits in learning from an adaptive condition rather than gains. We would have to consider factors such as: the poor

match between reaction times and learning strength, the possibility that learning strength does not inform scheduling strategy, the influence of participant strategies, the stochastic nature of memory representation such that fixed schedules somehow force order or otherwise smooth natural fluctuations in the learning process by distributing strength or attention across items, among other possibilities. We expect that some of these factors are real; that the strength of certain effects is sometimes diminished, that the learning process is more complex than as sometimes characterized, that there is noise in the behavioral process, and that we are still in an early stage of determining formal models of memory that can incorporate all of these minor, albeit influential, factors.

2.2 Experiment 1) Fixed Spacing vs. Adaptive Spacing

Brief description of study In order to explore issues relevant to the literature above, and to test whether adaptive spacing of delay intervals is more beneficial than fixed schedules of delay, we constructed an experiment that presented learners either adaptive or fixed schedules of practice under highly constrained learning conditions. In an initial session, participants learned factual items: 24 country names and locations on a map of Africa. Each item was presented exactly 4 times in all conditions. There were 3 different types of delay schedule: One group of participants received items using an adaptive algorithm as detailed in the explanation of adaptive sequencing in Chapter 1. Another group of participants received a fixed schedule of practice where half of their learning items were scheduled according to an equal schedule of practice (5-5-5 intervening items) and another half of their items was scheduled according to an expanding schedule of practice (1-5-9 intervening items). In the learning session, every presentation consisted of a test trial where participants were shown a country and asked to select a name. On every trial participants were told whether their responses were correct and were shown feedback containing the correct answer. Participants were measured first at a pre-test given before the learning session and at an immediate post-test given immediately after the learning session. The pre and post-tests were identical to

training trials except that there was no feedback given after a response. Each country was tested once in pre-test and once in post-test. Finally participants returned for a delayed post-test after one week. The delayed post-test was identical to the immediate post-test. If adaptive scheduling provides better learning conditions than fixed scheduling, participants should perform better on measures of recall at both immediate and delayed post-tests. It was expected that mean accuracies would be higher for items scheduled using an adaptive sequencing algorithm, since the pattern of spacing delays would be tuned to the interaction of learner, items and chance variability in learning events.

Planned analyses For this experiment, the primary dependent measures were mean accuracy and latencies (RT) across items. Since the total number of presentations for each item in each condition remains fixed, the analyses did not benefit from a comparison of learning efficiency (accuracy as a function of trials invested in learning). In addition to performance measures, the exact amount of spacing generated by adaptive schedules can be compared to those chosen for fixed schedules.

This experiment also served as a baseline for determining the individual item intervals for Expt. 2 (adaptive sequencing vs. fixed yoked schedules).

2.2.1 Methods

Participants The participants were 72 undergraduate psychology students who received course credit for completing the experiment.

Design There were two between-subject conditions, adaptive spacing and fixed spacing. There were two within subject fixed spacing conditions, equal fixed spacing and expanding fixed spacing. In the fixed spacing conditions, one random half of learning items was assigned to the fixed equal condition and the other half was assigned to the fixed expanding spacing condition.

Materials The learning materials consisted of 24 African countries that participants were required to identify on a map of Africa. 14 additional countries were used as 'filler' items in order to space presentations appropriately, especially at the end of a learning session (see note on filler items in 'Filler items and jitter in fixed schedules'). All material was presented on a computer within a web-based application. Participants saw a 500 pixel by 800 pixel map of Africa on the left side of the screen and a two column list of African countries alphabetically organized by column then row (see Figure 2.1). Each list label was a software button that could be independently selected using a computer mouse.

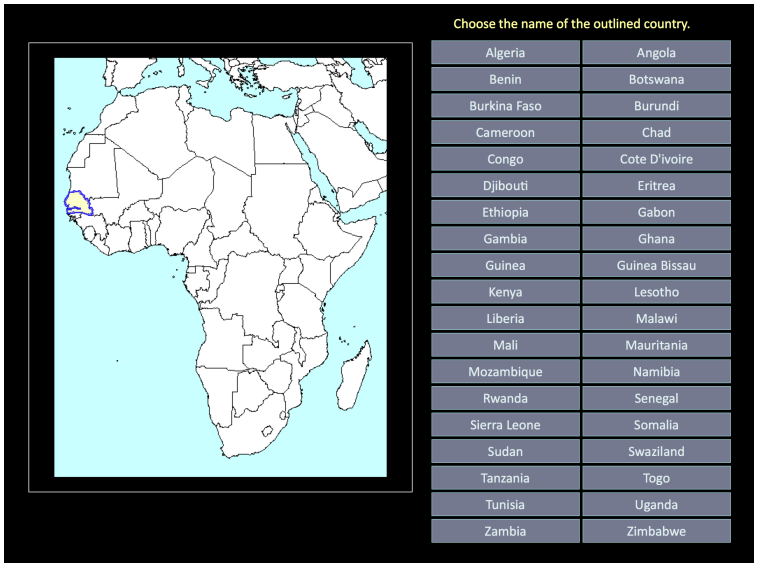

Figure 2.1: Exp 1. Stimuli presented to participants on each trial. Map of Africa with target country highlighted, and a list of response choices on the right side of screen.

Procedure In all sessions of the experiment learning items were presented singly, in the form of test trials. Participants were shown a map of Africa featuring an outlined country and were asked to select, from a list of labels containing country names, the name that matched the highlighted country. Participants used the computer mouse to select from the list of names.

Participants attended two sessions, separated by one week. In the first session, partici-

pants initially took a pre-test on all items. Pre-tests contained all 38 target and filler items, presented in random order. During the pre-test phase participants were not given feedback as to whether their response was correct. The pre-test was followed by a learning phase that consisted of the same type of trial as the pre-test, except that participants were given feedback after each response showing the correctness of their response as well as a label indicating the correct answer. The learning phase took up the majority of the first session of the experiment. During the learning phase, after every ten trials participants received block feedback indicating their average response accuracy and average response speed for the previous block of 10 trials and every previous block up to ten prior blocks. After the learning phase, an immediate post-test was administered, identical to that given in the pre-test. After the post-test participants were instructed to return in one week and were asked not to study or reflect on the information learned. A delayed post-test, identical to the immediate post-test, was administered after one week. No feedback was given on either post-test.

Spacing conditions Participants were randomly assigned to one of two scheduling conditions: fixed or adaptive (36 participants in each condition). In the adaptive condition, all training session trials were adaptively sequenced according to a reaction-time-based scheduling algorithm (see description of adaptive sequencing in Chapter 1). In the fixed condition, one random half of each participant's items were scheduled according to an equal spacing scheme, and the other random half were scheduled according to an expanding spacing scheme. Thus, in the fixed condition, every participant received two within-subject conditions that manipulated fixed scheduling in either an expanding or equal spacing scheme. This interleaving of conditions was done primarily to avoid the problem of excessive filler items in the expanding spacing condition.

In the fixed spacing condition, delays between presentations were pre-determined and constant. Items in the fixed equal condition received spacing of 5 trials between items. Items in the expanding spacing condition received first 1, then 5, then 9 trials between presentations of each item. For every participant in the fixed condition, the order of presentation was

pre-set so that every participant received the same number and order of fixed equal or expanding trials. Items in the fixed condition did not strictly alternate between equal and expanding schedules, but every attempt was made to balance the number of equal and fixed schedules across position in the entire learning phase, so as not to confound serial position with schedule type. Although the order of presentation of items in the fixed condition was fixed, the assignment of individual items to either of the two schedule types was randomized for each participant. In addition, the order of introduction of individual items was shuffled across possible positions in the pre-set schedule for each participant before the learning phase began.

For every participant in the adaptive condition, the total schedule order was dynamically decided during the learning session, and the order of introduction of new items was chosen randomly from the remaining items in a learning set for each participant.

In all conditions in Experiment 1, each learning item was presented a total of four times.

Filler items and jitter in fixed schedules Filler items were used to support the interleaving of items with different fixed schedules, and to maintain appropriate spacing delays at the end of a learning session, when no target learning items remained in the set. Filler items consisted of presentations of 14 additional countries, randomly selected whenever filler items were needed. Filler items are necessary in the fixed expanding presentation condition and the adaptive presentation condition; in both cases the final few presentations of items occurs at larger and larger delays, requiring filler items when no new target items are available. In addition to filler items at the end of learning, filler items are also required during the learning session. This is because the structure of fixed spacing intervals do not allow continuous presentation of items without conflicts in the intended interval schedule for each item. Imagine 4 items each presented first with a 1 trial delay, then a 3 trial delay. The layout of this presentation sequence would appear as follows: 1 2 1 2 3 4 1/3.. where the 7th presentation indicates a conflict between the first and third item. These conflicts do not arise with equal expanding schedules or with adaptive schedules. One of the strengths of

combining expanding and equal schedule presentations into the same session was that we were able to design a single fixed session that uses limited filler items for the majority of the learning session . We were able to do this by allowing for a degree of 'jitter' in any given fixed schedule following a simple rule: Each set of 3 delays between the four presentations of an item in the fixed or equal condition was allowed to deviate from it's pre-set interval (e.g., 1-5-9) by one position, smaller or larger at any delay except the first. For example, 1-6-9, 1-5-9, 1-5-10 would all be valid 'jittered' versions of the 1-5-9 expanding interval. Using jitter we were able to fit expanding and equal conditions together with minimal filler items. Thus, filler items were utilized primarily to fill expanding schedules at the end of training and their use was equated across both adaptive and fixed conditions.

Sequencing parameters The default adaptive sequencing parameters are described in Appendix, Table A.1. In this study, the default parameters were used for the adaptive algorithm except for the following parameters: 'RT weight', r = 3.0; 'enforced delay', D = 1.

2.2.2 Results

Average Accuracies across Test Phase Average accuracies for each experiment phase are presented in Figure 2.2 on page 41. At pre-test, adaptive accuracies were highest (M=0.076, SD=0.266), followed by fixed expanding (M=0.051, SD=0.22) and fixed equal (M=0.042, SD=0.20). Comparisons between conditions showed a reliable difference for adaptive vs. fixed equal ($t(70)=2.19$, $p=.032$), but not for adaptive vs. fixed expanding ($t(70)=1.54$, $p=.13$), or fixed equal vs. fixed expanding ($t(35)=0.73$, $p=.47$). These differences indicate some pre-test differences in performance across groups, despite random assignment of participants to conditions. Overall mean pre-test scores (M=0.056, SD=0.072) were significantly different from chance responding (one sample t-test: $t(107)=4.13$, $p<.01$), suggesting that some participants possessed prior knowledge of some countries. (Chance responding would have been 1 correct item out of 24, or .042.) As a result, we compute change scores between pre and post-test in addition to comparing average accuracies.

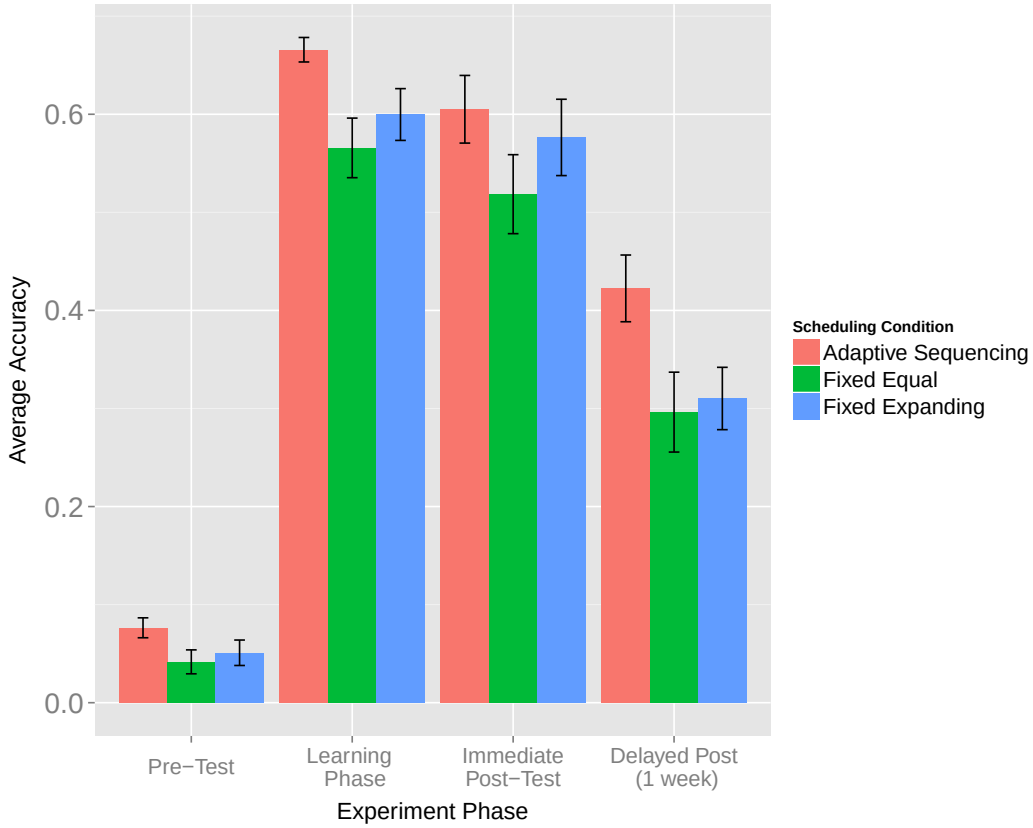

Figure 2.2: Exp 1. Average accuracies by experiment phase across 3 scheduling conditions. Error bars show +/- 1 standard error of the mean.

Performance results were not treated to a standard one-way ANOVA due to the special combination of between and within subjects factors. 3 ANOVAs were used, one for each comparison of pairs of conditions, with test-phase as a within subjects factor.

A 2X2 mixed factor ANOVA on scheduling condition (adaptive vs. fixed equal) and test phase (immediate post-test vs. delayed post-test) found a significant main effect of condition ($F(1,70)=4.63$, $p=.035$), a main effect of test phase ($F(1,70)=110.56$, $p<.001$) and no interaction of test phase and condition ($F(1,70)=1.06$, $p=.31$). A 2X2 mixed factor ANOVA on scheduling condition (adaptive vs. fixed expanding) and test phase found no reliable main effect of condition ($F(1,70)=2.37$, $p=.13$), a significant main effect of test phase ($F(1,70)=147.0$, $p<.001$), and a significant condition by test phase interaction ($F(1,70)=5.1$,

p=.027). A 2X2 repeated measures ANOVA on scheduling condition (fixed equal vs. fixed expanding) and test phase found a marginal main effect of condition ($F(1,70)=3.13$, $p=.081$), a main effect of test phase ($F(1,35)=126$, $p<.001$), and no condition by test phase interaction ($F(1,70)=1.18$, $p=.28$). A Bartlett's test confirmed homogeneity of variance for accuracies at both post-tests (immediate: $p=.64$, delayed: $p=.31$).

At the immediate post-test average accuracies were highest for the adaptive condition ($M=0.61$, $SD=0.21$), lower for the expanding condition ($M=0.58$, $SD=0.23$), and lowest for the equal condition ($M=0.52$, $SD=0.24$). Accuracies did not differ reliably between the adaptive and fixed conditions (adaptive vs. equal: $t(70)=1.63$, $p=.11$; adaptive vs. expanding: $t(70)=0.55$, $p=.58$). A paired t-test showed that the two within-subject fixed conditions differed reliably ($t(35)=2.15$, $p=.039$, Cohen's $d=0.24$). Looking at average accuracies at the delayed post-test, accuracies were highest in the adaptive sequencing condition ($M=0.42$, $SD=0.20$), and lower for the two fixed conditions: expanding spacing ($M=0.31$, $SD=0.19$) and equal spacing ($M=0.30$, $SD=0.24$). Individual comparisons showed average accuracies for the adaptive spacing condition were significantly greater than both of the fixed spacing conditions (adaptive vs. expanding: $t(70)=2.41$, $p=.019$, Cohen's $d=0.57$; adaptive vs. equal: $t(70)=2.38$, $p=.02$, Cohen's $d=0.56$). A paired t-test showed that the expanding and equal spacing means were not significantly different from each other ($t(35)=0.45$, $p=.65$).

Since there was detectable prior knowledge and marginally significant differences between conditions at pre-test, we examined post-test results in terms of change scores computed between pre and post-tests. We computed two types of change score for each participant: change scores and gain scores. Change scores were computed by subtracting average pre-test accuracies from average post-test accuracies. Gain scores were computed by subtracting pre-test scores from post-test scores, but only for those items that were accurate at both post-test and pre-test. Items that were answered accurately at pre-test but were answered inaccurately at post-test were not subtracted from post-test scores in the calculation of a gain score. This calculation is based on the assumption that there would not be a loss of knowledge between pre-test and post-test, since the primary intervention between tests was

a learning phase in which each item was presented with feedback.

Immediate post-test change scores were computed by subtracting a participant's average pre-test accuracy from their average post-test accuracy, and delayed post-test change scores were computed by subtracting average pre-test accuracy from average delayed post-test accuracy. Post-test and delayed post-test change scores are shown in Figure 2.3. Three ANOVAs

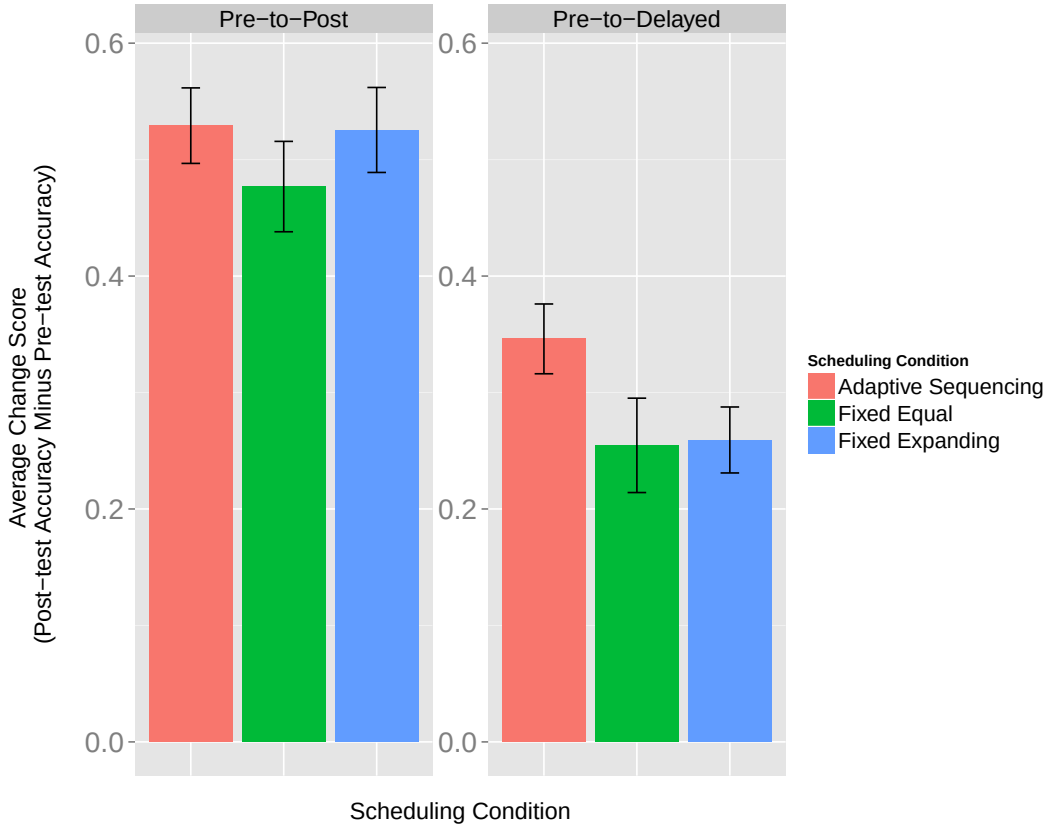

Figure 2.3: Exp 1. Average change in accuracy from pre to post-tests across 3 scheduling conditions. Left panel shows difference between immediate post-test and pre-test. Right panel shows difference between delayed post-test and pre-test. Error bars show +/- 1 standard error of the mean.

were conducted on change scores. A 2X2 mixed factor ANOVA on condition (adaptive vs. fixed expanding) and test phase found no main effect of condition ($F(1,70)=1.21$, $p=.28$), a

main effect of test phase ($F(1,70)=147.13$, $p<.001$), and a significant condition by test phase interaction ($F(1,70)=5.03$, $p=.028$). A 2X2 mixed factor ANOVA on condition (adaptive vs. fixed equal) and test phase found no main effect of condition ($F(1,70)=2.37$, $p=.13$), a main effect of test phase ($F(1,70)=110.68$, $p<.001$), and no significant condition by test phase interaction ($F(1,70)=1.03$, $p=.31$). A 2X2 repeated measures ANOVA on condition (fixed expanding vs. fixed equal) and test phase found no main effect of condition ($F(1,70)=1.29$, $p=.26$), a main effect of test phase ($F(1,35)=126$, $p<.001$), and no significant condition by test phase interaction ($F(1,70)=0.88$, $p=.35$).

At immediate test, changes scores were similar for all schedules and did not show significant differences (adaptive vs. fixed equal: $t(70)=1.03$, $p=.30$, Cohen's $d=0.22$; adaptive vs. fixed expanding: $t(70)=0.076$, $p=.94$, Cohen's $d=0.25$; paired t-test between fixed equal and expanding conditions: ($t(35)=1.58$, $p=.12$, Cohen's $d=0.22$). At delayed-test, average change scores were highest in the adaptive condition ($M=0.35$, $SD=0.18$), lowest in the fixed equal condition ($M=0.26$, $SD=0.24$) and nearly as low in the fixed expanding condition ($M=0.26$, $SD=0.17$). Comparing means at delayed test, t-tests showed a significant difference between the adaptive and fixed expanding condition ($t(70)=2.11$, $p=.04$, Cohen's $d=0.50$) and a marginally significant difference between the adaptive and fixed equal condition ($t(70)=1.81$, $p=.07$, Cohen's $d=0.43$). A paired t-test between the two fixed conditions showed no reliable difference (fixed equal vs. fixed expanding: $t(35)=0.13$, $p=.90$, Cohen's $d=0.02$).

In addition to change scores, we computed gain scores by subtracting pre-test scores from post-test scores, but did not include items that were accurate at pre-test but inaccurate at post-test. Gain scores are shown in Figure 2.4. Three ANOVAs were conducted on gain scores. An 2X2 mixed factor ANOVA on condition (adaptive vs. fixed expanding) and test phase showed no main effect of condition ($F(1,70)=1.67$, $p=.20$), a main effect of test phase ($F(1,70)=141.5$, $p<.001$) and a reliable condition by test phase interaction ($F(1,70)=5.51$, $p=.022$). An 2X2 mixed factor ANOVA on condition (adaptive vs. fixed equal) and test phase showed a marginal main effect of condition ($F(1,70)=3.09$, $p=.08$), a main effect of

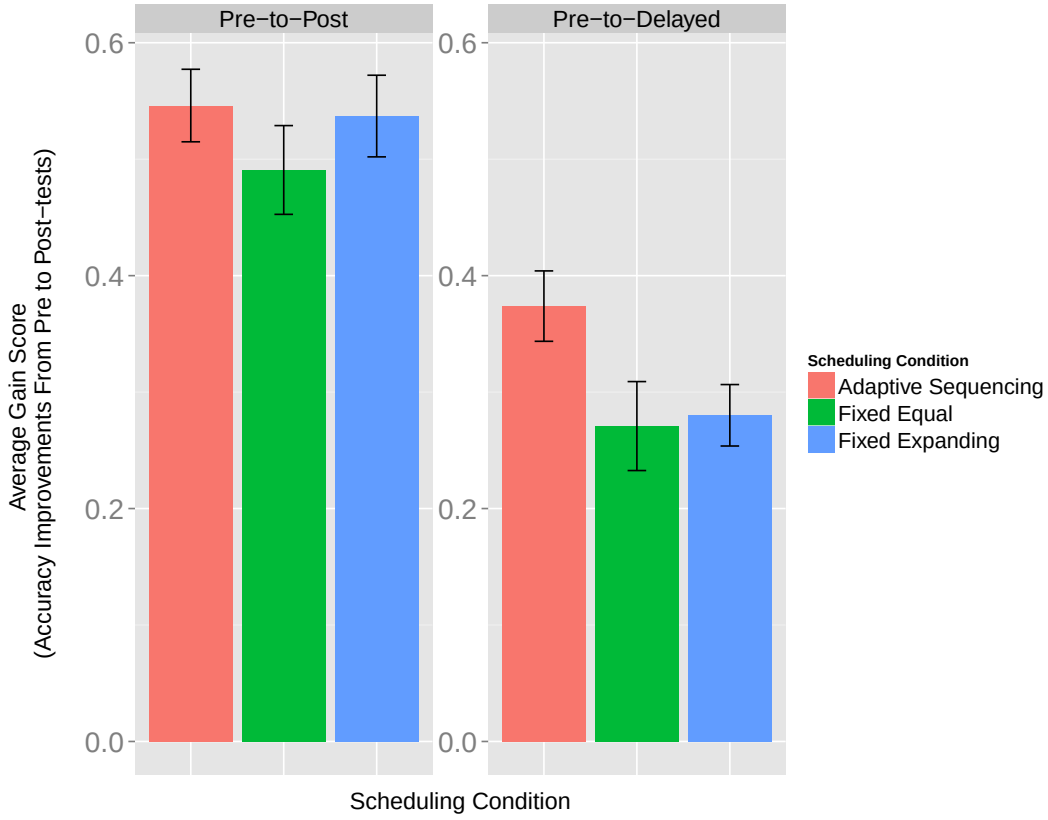

Figure 2.4: Exp 1. Average gain in accuracy from pre to post-tests across 3 scheduling conditions. Left panel shows difference between immediate post-test and pre-test. Right panel shows difference between delayed post-test and pre-test. Error bars show +/- 1 standard error of the mean.

test phase ($F(1,70)=105$, $p<.001$), and no condition by test phase interaction ($F(1,70)=1.55$, $p=.22$). An 2X2 repeated measures ANOVA on condition (fixed expanding vs. fixed equal) and test phase found no main effect of condition ($F(1,70)=1.89$, $p=.17$), a main effect of test phase ($F(1,35)=114$, $p<.001$), and no condition by test phase interaction ($F(1,70)=0.84$, $p=.36$).

At immediate test, mean change scores were similar (adaptive: $M=0.55$, $SD=0.19$; fixed equal: $M=0.49$, $SD=0.23$; fixed expanding: $M=0.54$, $SD=0.21$). Comparisons between conditions did not show significant differences: (adaptive vs. fixed equal ($t(70)=1.12$, $p=.26$,

Cohen's d=0.26; adaptive vs. fixed expanding: t(70)=0.19, p=.85, d=0.045; paired t-test between the two fixed conditions: t(35)=1.69, p=.10, Cohen's d=0.21). Comparing gain scores at delayed post-test, scores were highest for the adaptive condition (M=0.37, SD=0.18), and lower for fixed equal (M=0.27, SD=0.23) and fixed expanding schedule conditions (M=0.28, SD=0.16). Individual comparisons of conditions at delayed test showed significant differences between the adaptive and both fixed conditions (adaptive vs. fixed equal: t(70)=2.11 p=.01, d=0.50; adaptive vs. fixed expanding: t(70)=2.34, p=.02, Cohen's d=0.55). The two fixed spacing conditions did not differ reliably: t(35)=0.31, p=.76, Cohen's d=0.05). Since every condition showed significant differences between gain scores at immediate test and at delayed test (all ps<.001), the apparent interaction was driven by gain scores that were not significantly different at immediate post but were different at delayed test.

Average latencies (Reaction time) Average reaction times (RTs) are shown in figure 2.5 for each condition and for 3 experimental phases: the learning phase, immediate post-test and delayed post-test. RT data only include RTs from trials that were answered correctly. Pre-tests are ignored owing to the few items that were answered correctly in that phase. Of most interest were RTs at training and at immediate and delayed post-tests. ANOVAs were not run on RTs due to missing data for 4 participants who answered no items correctly at either immediate post-test or delayed post-tests. At training, adaptive RTs were lowest (M=4.04 sec, SD=0.99) followed by fixed equal (M=4.59, SD=1.34), then fixed expanding (M=4.61, SD=1.4). Individual comparisons showed the difference between the adaptive and the two fixed conditions was marginally significant (adaptive vs. fixed expanding: t(70)=1.97, p=.052, d=0.47; adaptive vs. fixed equal: t(70)=1.97, p=.053, d=0.47), but a paired t-test between the two fixed conditions showed no significant difference (t(35)=0.11, p=.17, d=0.012). At immediate post-test t-tests between conditions showed no significant difference between adaptive and the two fixed conditions (adaptive vs. fixed expanding: t(69)=0.07, p=.94, d=0.02; adaptive vs. fixed equal: t(70)=1.46, p=.15, d=0.35) and a paired t-test between the two fixed conditions showed no significant difference (t(34)=1.39,

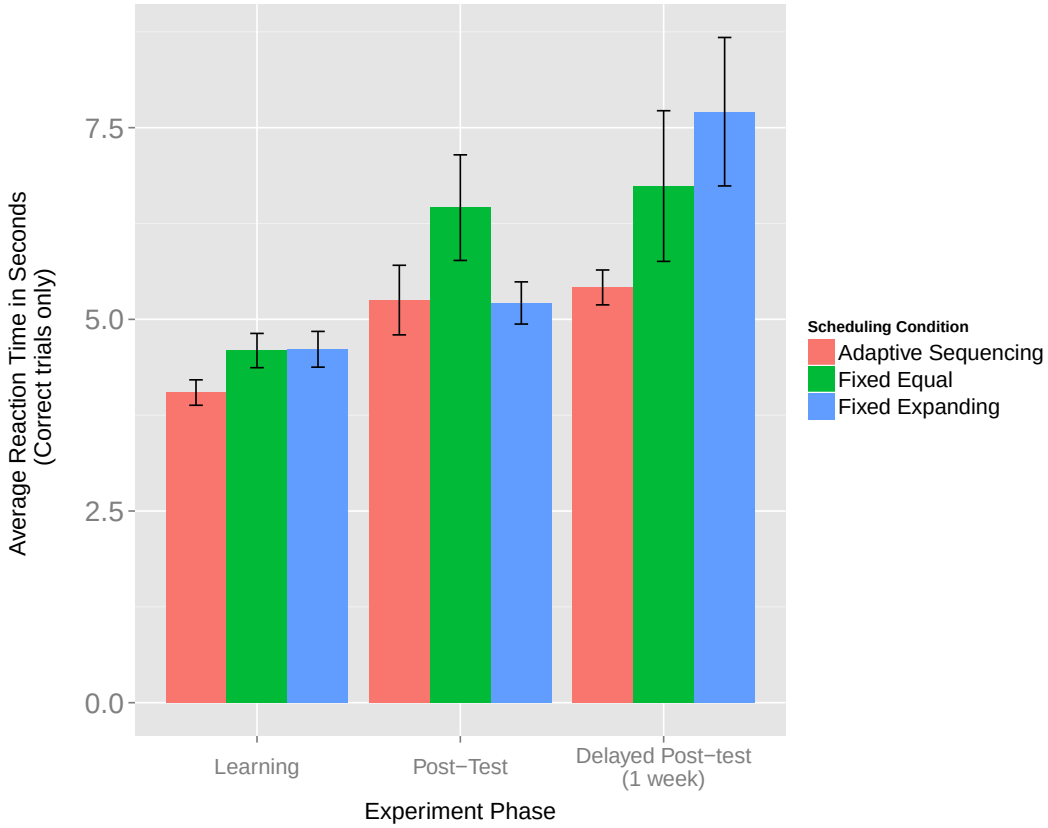

Figure 2.5: Exp 1. Average reaction time (in seconds) at each test phase across 3 scheduling conditions. Reaction time is from correctly answered trials only. Error bars show +/- 1 standard error of the mean.

p=.17, 0.43). At the delayed post-test there was a significant difference between the adaptive and fixed expanding conditions (t(69)=2.3, p=.02, d=0.64), but other RTs were not significantly different from one another (adaptive vs. equal: t(66)=1.31, p=.19, d=0.36; fixed equal vs. fixed expanding: t(30)=1.48, p=.15, d=0.166). Comparing RTs across post-test phases, only the difference between the fixed expanding condition at post-test vs. delayed post-test was significant (t(33)=2.8, p=.008; all other ps>.70).

We also examined the RTs at each presentation in learning across the 3 schedules. Reaction times during the learning phase are shown in Figure 2.6 by scheduling condition and presentation number. Examination of response times revealed that conditions did not differ

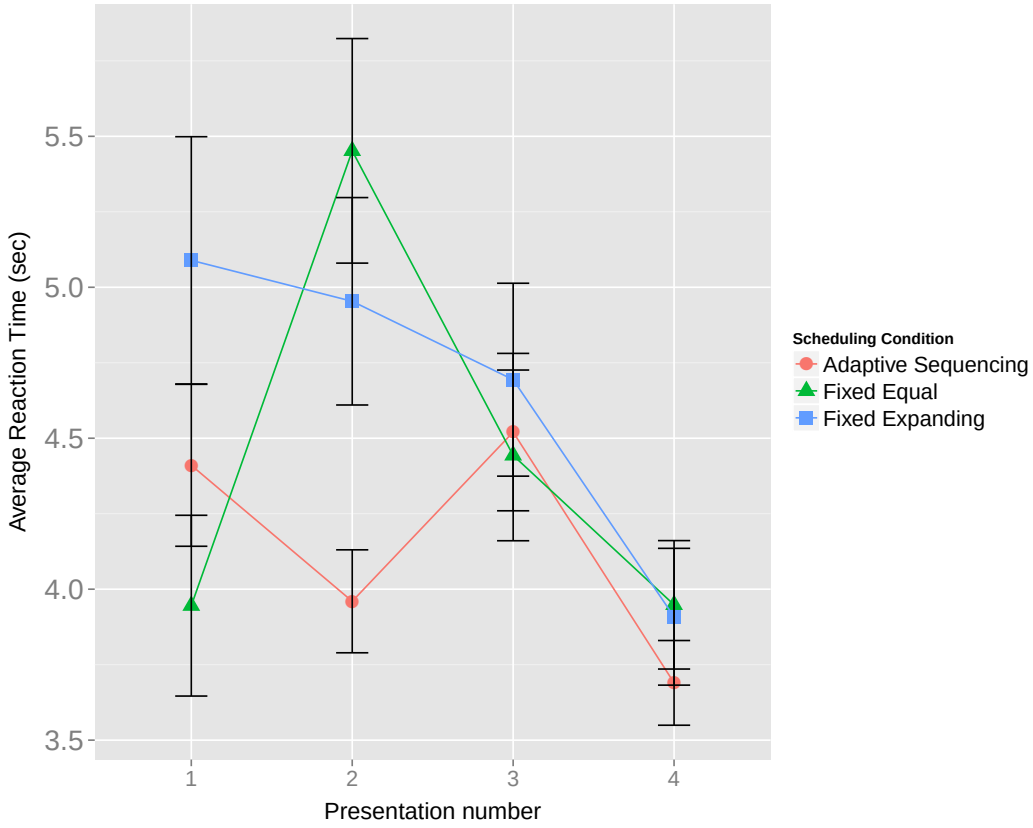

Figure 2.6: Exp 1. Average reaction time (in seconds) at each presentation (1-4) during learning, across the 3 scheduling conditions. Reaction time is from correctly answered trials only. Error bars show +/- 1 standard error of the mean.

in response times at the first, third, or fourth presentation (all t-test ps>.05), but that there were significant differences at the second presentation. There were significant differences between the adaptive and fixed conditions (adaptive vs. fixed expanding: $t(70)=2.59$, $p=.01$; adaptive vs. fixed equal: $t(70)=3.64$, $p<.001$) but not between the two fixed conditions ($t(34)=1.06$, $p=.29$, paired t-test).

Analyses of average delay size (Trials between presentations of an item) Adaptive and fixed conditions differed in the size of trial delays delivered to individual items during the learning session. The mean delay size per learner was calculated by averaging

the mean presentation delays for every item and averaging over items. In the equal and expanding fixed conditions, the intervals chosen in the experiment (1-5-9 and 5-5-5) ensured that mean delays were always 5 trials in length. Average adaptive schedule delays were close in length but with some variance (M=6.7, SD=2.033) [5]. We also looked at the size of delays conditional on whether the presentation before the delay was responded to correctly or not. The mean delay size by scheduling condition and conditional on response accuracy are shown in Figure 2.7. Although the mean adaptive delay size was similar to the mean

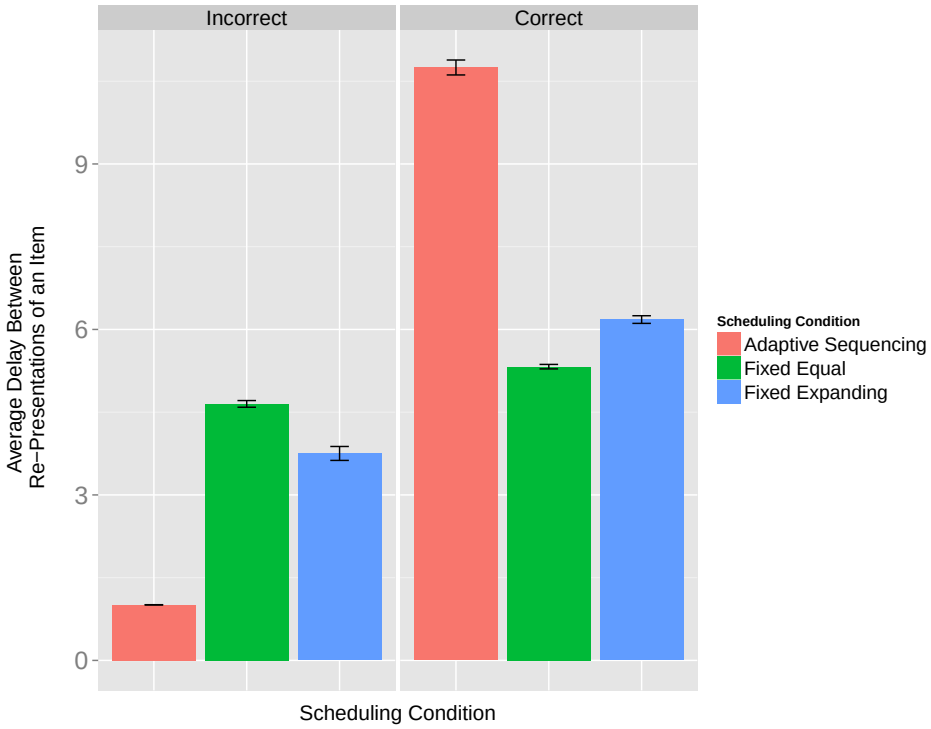

Figure 2.7: Exp 1. Average delay size (in trials) across 3 scheduling conditions conditional on whether the trial preceding the delay was answered correctly or not. Error bars show +/- 1 standard error of the mean.

delays for fixed schedules, the conditional values reveal that adaptive delays were both larger

[5] Across items, in the fixed conditions there was also some variance in delay size due to the use of 'jitter' and filler items

and smaller. After incorrect responses adaptive delays were small (M=1.01, SD=0.09) owing to the enforced delay mechanism. After correct responses adaptive delays were larger (M=10.88, SD=6.50) than average fixed delays.

Finally we examined the average delay at each presentation number for the adaptive condition. The mean sizes of the three delays in the adaptive condition are shown in Figure 2.8. The mean initial delay was the smallest (M=1.62, SD=0.71), the second delay largest (M=10.95, SD=4.46), and the third delay smaller than the second delay (M=7.52, SD=2.20). While it appears that the pattern of retrievals was not expanding, but expanding-then-

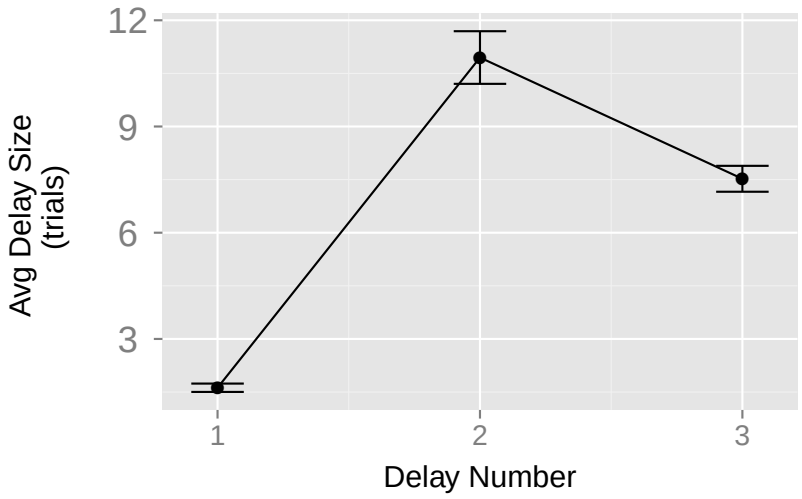

Figure 2.8: Exp 1. Average delay size (in trials) across 3 delays in the adaptive scheduling condition. Error bars show +/- 1 standard error of the mean.

contracting, in fact, a line of best fit to these points still yields a positive slope. There were also 4 adaptive participants who had strictly expanding practice - positively increasing delay sizes at each presentation.

2.2.3 Discussion

Learning was greater – as demonstrated by a variety of measures – for items scheduled using an adaptive sequencing algorithm than for items scheduled using pre-determined, fixed schedules of practice. Because average pre-test knowledge was slightly greater than chance and because there were slight differences between participants' prior knowledge across conditions, we relied on post-test accuracy as well as two derived measures of accuracy that discounted prior knowledge from measures of learning. 'Change scores' were computed by subtracting pre-test accuracy from post-test accuracy for each participant. 'Gain scores' were computed by subtracting from post-test accuracy only those items that were known at both pre-test and post-test. Both measures showed significant differences in learning improvements across scheduling conditions. Learners experienced significantly greater learning in the adaptive condition than in the equal fixed condition as measured by a change score – and when measured with a gain score, learners showed significantly stronger gains in the adaptive scheduling condition than in either of the fixed scheduling conditions. In addition, these improvements were retained across a considerable delay (1 week) suggesting that adaptive scheduling techniques produce not only greater learning, but more robust learning – that is, learning is more durable and long-lasting.

In addition to learning gains measured by accuracy, there appeared to be greater fluency - that is, faster responding - for participants who learned using an adaptive scheduling algorithm than for participants who learned using fixed expanding schedules of practice, suggesting that adaptive scheduling produces not only greater retention of knowledge, but deeper and stronger knowledge gains. Fluency is an important learning goal that is a marker of higher levels of expertise in many domains (especially domains that involve perceptual learning; Kellman & Garrigan, 2008).

There are two primary conclusions we can draw from these results. First, as relates to some debates in research on learning and memory, it appears that certain fixed schedules of practice are generally poorer at promoting learning than schedules generated using an

adaptive spacing algorithm. Second, between some select fixed schedules that we have tested, we can determine that expanding practice in this learning context is generally no better than equal practice. Finally we show that the patterns of delays associated with adaptive schedules were on the whole of a type that increases spacing delays across the duration of learning - although spacing tended to flatten out and contract at the last interval of practice.

Why were expanding vs. equal spacing schedules nearly the same? We can speculate as to why expanding intervals of practice were no different from equal intervals at a delayed test. First, we chose to keep those schedules as within-subject conditions (primarily to limit the amount of filler items that were needed to support the successful interleaving of items from expanding schedules). Combining conditions in this way may have influenced the overall learning benefits of either schedule - for example by reducing the power of any prediction or encoding strategies based on one particular kind of recurring pattern in a schedule. This null difference between the conditions might also be similar to null effects in the literature previously discussed. Given the particular learning items we used, the particular learners tested, and the particular pattern of learning strength each item possessed during a session, there may have been deficits to both expanding and to equal schedules, equating their overall performance. In particular, it is possible that larger expanded intervals could have been beneficial. However, as we found expanding patterns to be particularly successful in the adaptive condition, and as we saw that expanding schedules resulted in greater raw accuracy at an immediate post-test, our results do not contradict the idea that expanding retrieval practice is often an effective arrangement for learning.

Why did adaptive not do better? We can also speculate as to some possible reasons that adaptive sequencing did not outperform fixed schedules by a greater margin than was shown. Some reasons focus on potential benefits of fixed schedules. It is possible that some patterns of fixed scheduling may be superior for encouraging encoding strategies that focus learners' attention. For example, expectation of the recurrence of certain items may prime learners to better encode items during future occurrences. It is also possible that

the relatively small amount of practice with items during a short, fixed schedule tipped off learners as to the importance of having to study items relatively quickly. These sorts of schedule-mediated encoding decisions could probably be eliminated given new experiments designed to study or remove such effects.

Other reasons focus on the potential limitations of the adaptive scheduling parameters selected for use here. It is possible (even probable) that adaptive schedules were simply not as effective as they could have been. Some evidence from reaction time data (Figure 2.6) appear to indicate that reaction times were lower for the adaptive condition, relative to fixed schedules. This was especially true at presentation 2, and even at presentation 4. Low RTs reflect the fact that item recalls were not as difficult as they could have been. Combined with data that shows adaptive schedules tended to have extremely large delays between presentation 2 and 3, but somewhat shorter delays between 3 and 4, we can only suspect that participants in adaptive schedules received a somewhat suboptimal pattern of trial delays. Contracting spacing indicates a potential 'overshoot' of prior delay intervals, possibly due to misinterpretation of fast reaction time data after initial enforced delays. In the adaptive condition as implemented here, when an item is answered incorrectly, it recurs after 1-2 trials. This rapid recurrence (after an enforced delay) is meant to ensure that learning strength gets a foothold. However, the learner's response when the item recurs after 1-2 trials may reflect both long-term learning strength and some residual effects of the recent presentation. This kind of problem is the basis for separating two notions of learning strength in earlier work (Bjork & Bjork, 1992). In the present context, it may be that response time on the first successful trial after a miss should be adjusted somehow to reflect the residual effects of the recent trial (and feedback). Alternatively, the recurrence interval after a miss, with feedback provided, could be increased, albeit at the risk of further incorrect responses for difficult-to-learn items. These issues in adaptive scheduling pose interesting questions for future work.

Another general reason that effects were not as strong as hoped has to do with the conditions of learning. In general, scheduling conditions constrained the total amount of

learning; each item was presented a limited number of times (four) and in relation to the total amount of learning time, delayed post-tests came at a considerable delay (one week). It is encouraging that despite those limitations the effects were still visible and robust.

It is important, when analyzing these results and reflecting on their importance, to balance consideration of the available sources of evidence, including null-hypothesis statistical results, statistical power, observed effect sizes, and common-sense knowledge of the properties of the quantities measured. Although there were a number of marginally significant null-hypothesis tests, every critical condition comparison showed a medium effect size. These issues speak to the difficulty of obtaining reliable measures of differences when participants learn under highly artificial experimental conditions.

Possible criticisms We can identify some possible criticisms of the demonstrations in this experiment. One potential problem is the comparability of intervals between adaptive and fixed conditions. If intervals were longer on average in the adaptive case, greater spacing alone may have led to greater learning benefits in the adaptive condition. Other objections point at the choice of fixed delays. Those delays were somewhat arbitrarily chosen from the research on spacing effects and may have been incompatible with the type of stimuli chosen for the current study. Another criticism is that, since intervals were more varied in the adaptive case, it was some aspect of the variable contexts of encoding that promoted learning, not anything about delay size or character.

These are interesting issues, but they do not detract from the major claims of this study and further experiments do much to dispel these questions. In Experiment 2, where we attempt to find the locus of learning effects in adaptive schedules, we compare adaptive schedules with fixed schedules that are matched to both the character of the stimulus set and that have similar patterns of spacing delays to that generated by adaptive schedules. In experiments in Chapter 3 we test schedules that have greater amounts of variability across schedule delays.

Possible explanations for effects There are a couple of immediate explanations for the results demonstrated in this experiment. First is that adaptive delays generated schedules that better matched the ongoing learning dynamics of individual items and learners. This includes such characteristics as the varying difficulty of individual items or the varying abilities of individual learners. In addition, events such as unexpected variation or fluctuation in the learning strength of items would likely be smoothed out by an adaptive spacing schedule. The variable characteristics of learners and items implies that spacing intervals for those items should also vary. Because fixed schedules did not vary as a function of either items or learners, performance suffered. Adaptive schedules appeared to detect and respond to variation in learning to a significant degree. The result was better post-test and delayed post-test retention, as well as better speed of responding, for participants who learned using an adaptive scheduling algorithm than those who learned using some select fixed schedules of practice. This experiment is the first to show that adaptive schedules perform better when compared directly with fixed schedules of practice.

Experiment 1 Conclusions Despite these potential concerns and despite the modest nature of some of the learning outcomes, this study reflects importantly on the primary hypotheses. It also adds useful knowledge to continuing, complex debates about the efficacy of schedules of practice in the research literature. Experiment 1 gave strong evidence that adaptive sequencing provides efficacy beyond some common fixed spacing schedules that are known to improve learning in many domains.

2.3 Experiment 2) Yoked-adaptive Fixed Spacing vs. Adaptive Sequencing.

To what can we attribute the learning improvements exhibited by adaptive scheduling using an adaptive sequencing algorithm? How can we identify the driver of those improvements? So far we have suggested that the power of adaptive intervals rests on adaptation to ongoing

learning strength during learning. We have offered two possible sources of this advantage: adaptation to individual items and adaptation to individual learners. In order to more fully explore these possible effects, we designed a set of fixed schedules that either targeted or ignored individual item differences during learning. We compared adaptive spacing to fixed delay intervals where the fixed schedules had spacing intervals that were copied directly from adaptively generated schedules. Fixed schedules in Experiment 2 were 'yoked' to the exact same schedule characteristics as adaptive schedules. These new 'yoked' participants thus received schedules that identically mimicked the size of spacing delays and order of item presentation from prior adaptive schedules, but presentations were strictly fixed - that is they did not change depending upon a participant's ongoing responses or as any direct function of ongoing learning strength. Across two fixed conditions, schedules were either yoked to preserve individual item characteristics, or item characteristics were ignored by shuffling items across delays. In the former case, fixed schedules compared to adaptive schedules would demonstrate the power of individual learners. In the latter case – fixed schedules with item characteristics ignored compared to fixed schedules with item characteristics preserved – the comparison would demonstrate the power of adaptation to item characteristics. These comparisons would help unveil the power of adaptive schedules of practice in improving long-term learning.

This experiment utilized the same learning material as Experiment 1 and retained the pre-test, post-test, delayed post-test design, but changed the nature of the fixed condition. As a side effect, these fixed schedules were expected to be more competitive with adaptive schedules since they were tuned to a length roughly appropriate for the type of learning material and more closely matched to that of adaptive schedules than in Experiment 1.

Yoking fixed schedules to adaptive schedules was accomplished in the following way: a participant in a 'yoked' condition received the same schedule of delays that a prior participant in an adaptive condition received. The difference between the two schedules was that in the yoked conditions, spacing delays were fixed, and the schedule had no relation to the 'yoked' participants' ongoing pattern of performance during learning. These fixed, yoked

schedules were copied directly from the trial record of actual participants who had recently completed the adaptive condition in the same ongoing experiment. Because yoked schedules deliver spacing delays that have been adaptively generated as a function of prior learners' performances with the same material, they represent one version of delays that have been optimized or tuned for the duration and character of the current learning stimuli[6]. In this way 'yoked' schedules served as a direct comparison between adaptive schedules and fixed schedules. In addition, two separate 'yoked' conditions were designed to tease apart learner vs. item characteristics. In a yoked-item condition, a learner received the same schedule of delays that a prior adaptive participant received: each item was presented in the same order, and the pattern of delays given to each item was retained. In a yoked-random condition, a learner received the same schedule of delays that a prior participant received, but items were shuffled across the pre-specified schedule of delays. In other words, if the prior adaptive learner received the country 'Angola' with a 1-5-15 series of delays, a yoked-item user would get the same item, at the same point in the learning session with the same delays. A yoked-random learner on the other hand, would receive the same series of delays (1-5-15) but for a different item (e.g., 'Botswana').

The purpose of this manipulation was two-fold: First, the difference between adaptive and yoked-item schedules distinguishes between learning schedules which do or do not take into account individual learner differences. In the adaptive case, schedules take into account learner differences, in the yoked-item condition schedules ignore learner differences, but - to some degree - preserve item differences. Second, the difference between yoked-item and yoked-random conditions explains the effect of schedules that do or do not take into account item differences during learning (neither takes into account learner differences).

[6]Another type of optimization might include averaging across all item delays for participants in an adaptive condition and yoking new participants' delays to these averages. Of many possible variations along these lines we chose direct yoking because of the power it gave us in generating subconditions that examine aspects of item vs. learner differences in learning.

2.3.1 Method

Participants The participants were 48 undergraduate psychology students who received course credit for completing the experiment.

Materials The learning materials were identical to Experiment 1, that is, 24 African countries as well as filler items.

Design This experiment examined a type of 'yoked' fixed spacing in relation to adaptive spacing. In three between-subject conditions (16 participants per condition), learning items were presented to participants in either an adaptive schedule (identical to Experiment 1), or in one of two 'yoked' fixed schedules. Each of the fixed conditions were assigned a single adaptive yoked 'target' participant, usually the participant who had last run in the adaptive condition. In every condition schedules were designed to present each item four times.

Procedure The order of the pre-test, learning phase, post-test and delayed post-tests were identical to Experiment 1. Trial presentations were identical to Experiment 1.

Yoking conditions For every participant in the adaptive condition, the total schedule order was dynamically decided as the participant completed the learning session, and the order of introduction of new items was chosen randomly from the remaining items in a learning set for each participant. For participants in either of the yoked conditions, items were presented on a fixed, pre-set schedule. The schedule was based on a prior adaptive participant's trial record. In the yoked-item condition, the trial record was simply copied, so that a new yoked participant received a duplicate version of the trial record of the prior adaptive participant including the order of introduction of items, the size of the spacing delays delivered to items, and the number and schedule of filler items. In a yoked-random condition, the trial record of the previous adaptive participant was retained but the assignment of items was shuffled, so that new yoked participants received the same schedule of delays, but the

set of items that received those delays was shuffled. For example, if a prior participant in the adaptive scheduling condition received a series of three delays, 2-4-10 for the item 'Angola', a participant in the yoked-item condition would get the same item, at the same point in the learning session with the same delays, whereas a participant in the yoked-random condition would receive the same series of delays but for a different item (e.g., 'Botswana'). For each pair of target adaptive and yoked fixed conditions, then, every series of delays (e.g., 2-4-10) occupied the same serial position in the learning session, but for the yoked-random group, items were shuffled across the pre-set series of delays. It should also be noted that not every set of yoked conditions had a unique yoked target adaptive condition. This was due to occasional errors with participant login to the system. The most common result was the occasional re-use of the last most recent adaptive condition as a target for a new set of yoked participants.

2.3.1.1 Expected results

It was expected that learners in the adaptive scheduling condition would fare better than learners in either yoked condition. Between the two yoked conditions, it was expected that yoked-item learners would perform better than those in the yoked-random condition, since yoked-item schedules would better match the pattern of increases in learning strength for individual items than would the yoked-random condition. In addition, if the yoked-item condition performance was roughly equal to adaptive performance, but better than yoked-random performance this would indicate the general benefit of tuning patterns of delay to schedules that approximate ideal patterns for individual items. If however, performance was equal across all three conditions, the supposed advantage of adaptive over fixed schedules demonstrated in Experiment 1 would likely have to be due to specific aspects of the fixed schedules used in Experiment 1, not a general advantage over fixed schedules.

2.3.2 Results

Average accuracies for each experiment phase are presented in Figure 2.9.

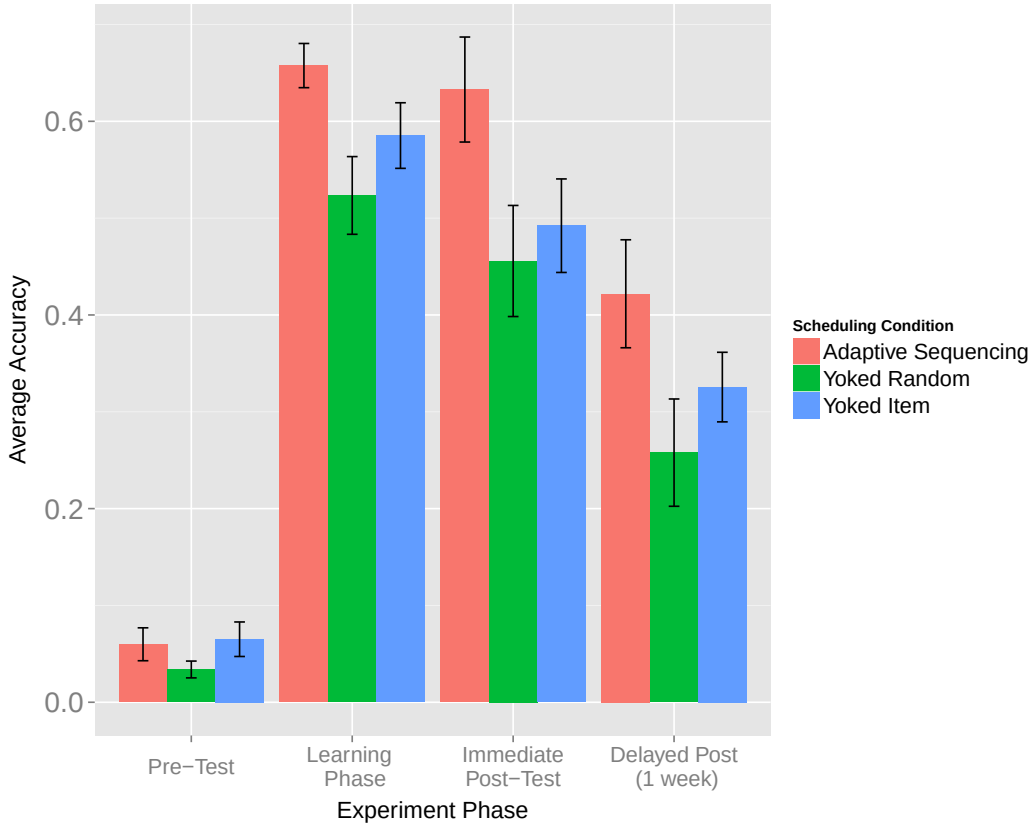

Figure 2.9: Exp 2. Average accuracies by experiment phase across 3 scheduling conditions. Error bars show +/- 1 standard error of the mean.

Pre-test scores At pre-test, an ANOVA did not find significant differences between scheduling conditions ($F(2,45)=1.23$, $p=.30$). Overall mean pre-test scores ($M=0.053$, $SD=0.061$) were significantly different from chance responding (one sample t-test: $t(47)=2.88$, $p<.01$), suggesting that some participants possessed prior knowledge of some countries. As a result, we compute change scores between pre and post-test in addition to comparing average accuracies.

Average post-test accuracies Analyzing accuracies at immediate and delayed post-test, a 3X2 mixed factor ANOVA on condition (adaptive vs. yoked-random vs. yoked-item) and test phase (immediate post-test vs. delayed post-test) found a significant main effect of condition ($F(2,45)=3.3$, $p=.046$), a significant main effect of test phase ($F(1,45)=77.09$, $p<.001$), and no condition by test phase interaction ($F(2,45)=0.36$, $p=.7$). A Bartlett's test confirmed homogeneity of variance for accuracies at both post-tests (immediate: $p=.80$, delayed: $p=.19$). At the immediate post-test, average accuracies were highest for the adaptive condition ($M=0.63$, $SD=0.22$), lower for the yoked-item condition ($M=0.49$, $SD=0.19$), and lowest for the yoked-random condition ($M=0.46$, $SD=0.23$). Comparing means at the immediate post-test, t-tests showed average accuracies for the adaptive spacing condition were reliably greater than the yoked-random condition ($t(30)=2.24$, $p=.03$, Cohen's $d=0.80$) and adaptive spacing marginally exceeded the yoked-item condition ($t(30)=1.94$, $p=.062$, Cohen's $d=0.69$). The two yoked conditions did not differ from one another ($t(30)=0.49$, $p=.63$, Cohen's $d=0.17$). Accuracies at the delayed post-test were highest in the adaptive sequencing condition ($M=0.42$, $SD=0.22$), lower for the yoked-item condition ($M=0.326$, $SD=0.144$) and lowest for the yoked-random condition ($M=0.26$, $SD=0.22$). Similar to the immediate post-test, at the delayed post-test, average accuracies for the adaptive spacing condition were reliably greater than the yoked-random condition ($t(30)=2.09$, $p=.045$, Cohen's $d=0.74$), but did not reliably exceed the yoked-item condition ($t(30)=1.45$, $p=.16$, Cohen's $d=0.53$). The two yoked conditions did not differ ($t(30)=1.03$, $p=.31$, Cohen's $d=0.37$).

Change and gain scores Since there was measurable prior knowledge we examined post-test results in terms of change scores computed between pre and post-tests. We computed the same two types of change score as Experiment 1: change scores and gain scores. Change scores were computed by subtracting average pre-test accuracies from average post-test accuracies. Gain scores were computed by subtracting pre-test scores from post-test scores, but only for those items that were accurate at both post-test and pre-test.

Immediate post-test change scores and delayed post-test change scores are shown in

Figure 2.10. Analyzing change scores at immediate and delayed post-test, a 3X2 mixed

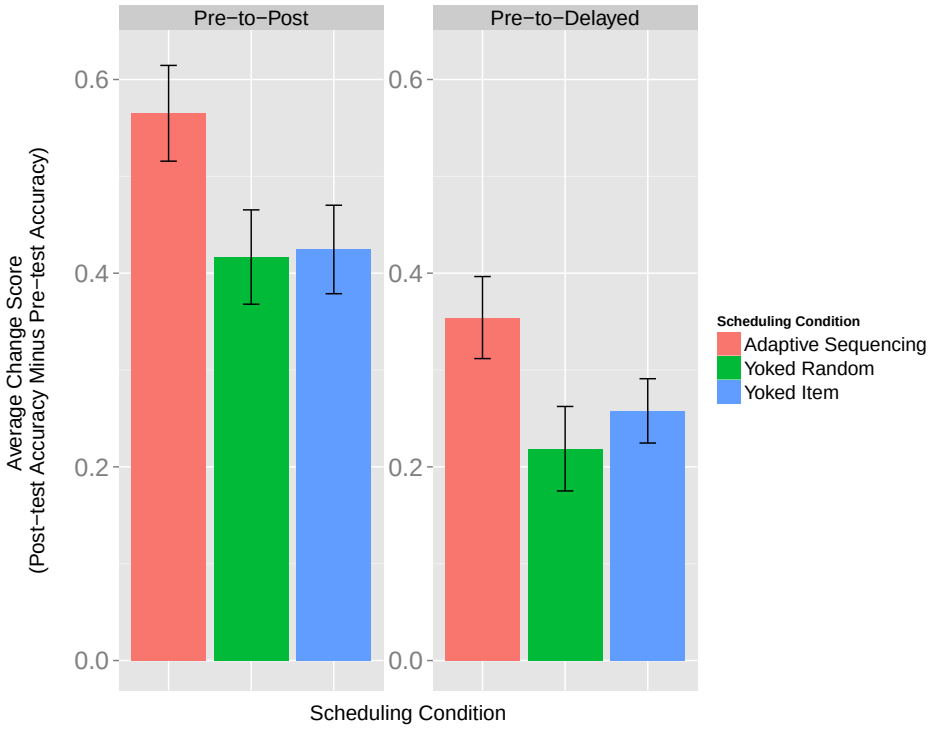

Figure 2.10: Exp 2. Average change score at immediate and delayed post-tests. Error bars show +/- 1 standard error of the mean.

factor ANOVA on condition and test phase found a significant main effect of condition ($F(2,45)=3.64$, $p=.034$), a significant main effect of test phase ($F(1,45)=77.09$, $p<.001$), and no condition by test phase interaction ($F(2,45)=0.36$, $p=.7$). Change scores at immediate post-test were highest in the adaptive condition ($M=0.57$, $SD=0.20$), lowest in the yoked-random condition ($M=0.42$, $SD=0.20$) and nearly as low in the yoked-item condition ($M=0.42$, $SD=0.18$). Comparing means, t-tests were significantly different between the adaptive and both of the two yoked conditions (adaptive vs. yoked-item: $t(30)=2.09$, $p=.045$, Cohen's $d=0.74$; adaptive vs. yoked-random: $t(30)=2.14$, $p=.04$, Cohen's $d=0.76$) but not a significant difference between the two yoked conditions ($t(30)=0.12$, $p=.91$, Cohen's $d=0.04$). Delayed post-test change scores were lower but similar: average scores were highest in the

adaptive condition (M=0.35, SD=0.17), lowest in the yoked-random condition (M=0.22, SD=0.17) and nearly as low in the yoked-item condition (M=0.26, SD=0.13). Comparing means, t-tests showed significant differences between the adaptive and the yoked-random conditions (t(30)=2.23, p=.03, Cohen's d=0.78), and a marginally significant difference between adaptive and yoked-item (t(30)=1.79, p=.08, Cohen's d=0.63). The difference between the two yoked conditions was not significant (t(30)=0.71, p=.48, Cohen's d=0.25).

In addition to change scores, we also computed gain scores by subtracting pre-test scores from post-test scores, but only for those items that were accurate at both post-test and pre-test. Items accurate at pre-test but inaccurate at post-test were not subtracted from post-test scores. Gain scores are shown in Figure 2.11. A 3X2 mixed factor ANOVA on condition and

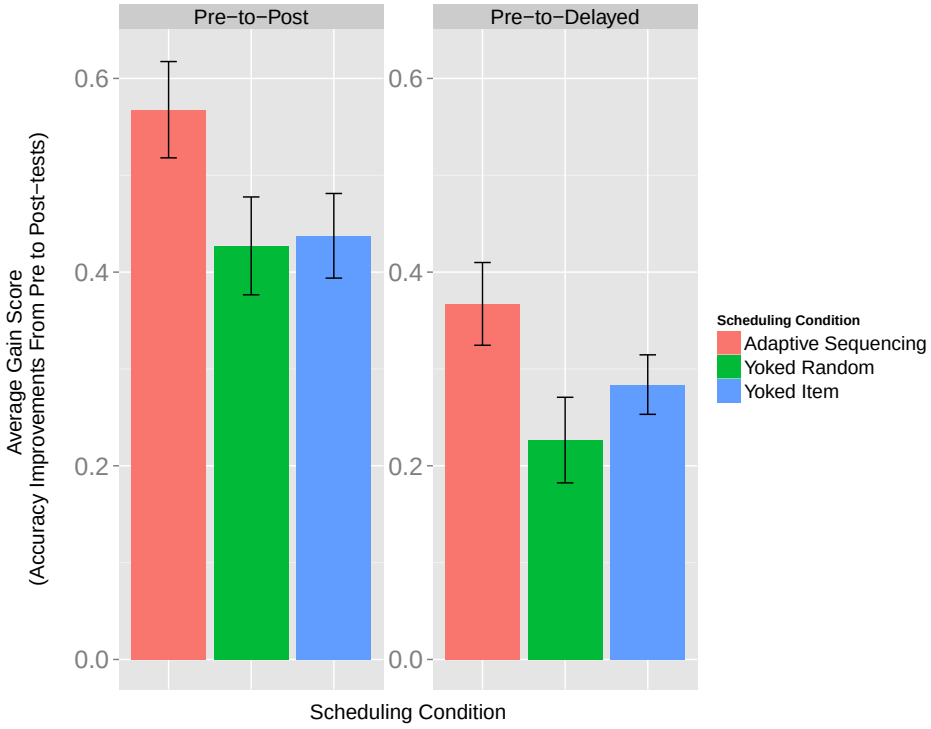

Figure 2.11: Exp 2. Average gain scores at immediate and delayed post-tests. Error bars show +/- 1 standard error of the mean.

test phase found a significant main effect of condition (F(2,45)=3.3, p=.046), a significant

main effect of test phase ($F(1,45)=77.09$, $p<.001$), and no condition by test phase interaction ($F(2,45)=0.36$, $p=.70$). Like change scores, gain scores at delayed post were highest for the adaptive condition (M=0.367, SD=0.171), lower for the yoked-item condition (M=0.284, SD=0.123) and lowest for the yoked-random schedule condition (M=0.227, SD=0.177). Comparing gain scores by condition at immediate post-test, t-tests showed marginally significant differences between the adaptive and yoked-random conditions ($t(30)=1.98$, $p=.057$, Cohen's $d=0.70$), between adaptive and the yoked-item condition ($t(30)=1.97$, $p=.058$ Cohen's $d=0.70$) and no difference between the two yoked conditions ($t(30)=0.16$, $p=.88$, Cohen's $d=0.06$). Comparing gain scores by condition at delayed post-test, t-tests showed significant differences between the adaptive and yoked-random conditions ($t(30)=2.29$, $p=.03$, Cohen's $d=0.81$) but not between adaptive and the yoked-item condition ($t(30)=1.58$, $p=.12$, Cohen's $d=0.57$) or between the two yoked conditions ($t(30)=1.06$, $p=.30$, Cohen's $d=0.38$).

Average latencies (Reaction time) Average reaction times (RTs) are shown in Figure 2.12 for each condition and each experimental phase except pre-tests (pre-tests are ignored owing to the few items that were answered correctly in that phase). RT data only include RTs from trials that were answered correctly. Of most interest were RTs at training and at immediate and delayed post-tests. At the learning phase, an ANOVA showed no significant differences between conditions ($p>.18$). A 3x2 ANOVA on scheduling condition and the two post-test phases found no significant effect of condition ($F(2,45)=0.26$, $p=.77$), no effect of test phase ($F(1,45)=1.12$, $p=.29$), and no interaction of scheduling condition with post-test phase ($F(2,45)=0.03$, $p=.97$). Paired comparisons showed that RTs at each phase were not significantly different from one another (all $ps>.05$). Comparing RTs across post-test phases, no conditions showed significantly different RTs across post-test phase (all $ps>.05$).

We also examined the RTs at each presentation in learning across the 3 schedules. Reaction times during the learning phase are shown in Figure 2.13 by scheduling condition and presentation number. A mixed factor, 3X4 ANOVA on scheduling condition and presentation number (1-4) found significant main effects of presentation number ($F(3,42)=4.93$, $p<.001$),

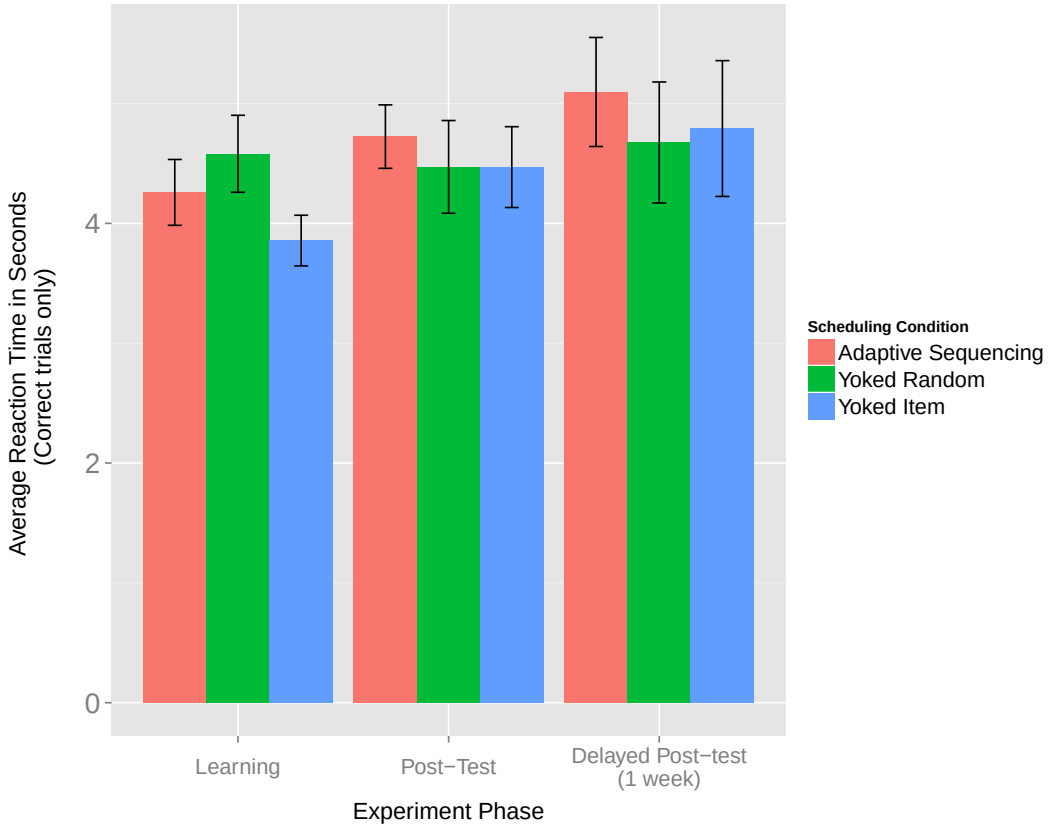

Figure 2.12: Exp 2. Average reaction time (in seconds) at each test phase across 3 scheduling conditions. Reaction time is from correctly answered trials only. Error bars show +/- 1 standard error of the mean.

no effect of condition ($F(2,42)=0.52$, $p=.59$), and no significant presentation number by condition interaction ($F(6,115)=0.61$, $p=.72$). Examination of response times revealed that conditions did not differ in response times at any presentation (all t-test $ps>.05$). Comparing the difference between presentation numbers for each condition, there were no significant differences between any presentation number except for adaptive between presentation 2 and 3 ($t(15)=2.14$, $p=.049$) and 3 and 4 ($t(15)=3.7$, $p=.002$), and yoked-item between 2 and 3 ($t(15)=3.13$, $p=.007$) and 3 and 4 ($t(15)=2.29$, $p=.04$) (all other $ps>.05$).

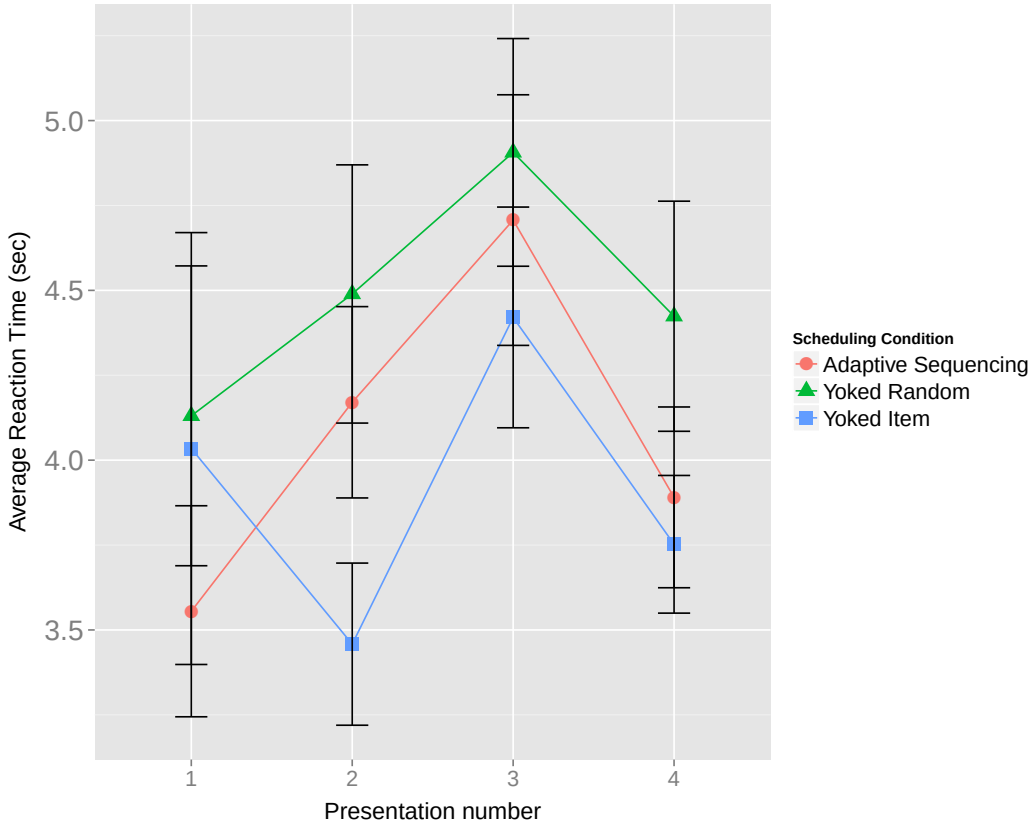

Figure 2.13: Exp 2. Average reaction time (in seconds) at each presentation (1-4) during learning, across the 3 scheduling conditions. Reaction time is from correctly answered trials only. Error bars show +/- 1 standard error of the mean.

Analyses of average delay size (Trials between presentations of an item) Adaptive and yoked conditions differed only slightly in the size of trial delays delivered to individual items during the learning session. The mean delay size per learner was calculated by averaging the mean presentation delays for every item and averaging over items. Average adaptive schedule delays were close in length to the adaptive condition in Experiment 1 (adaptive: M=6.77, SD=1.46, yoked-item: M=6.81, SD=1.40, yoked-random: M=6.89, SD=1.43). We also looked at the size of delays conditional on whether the presentation before the delay was responded to correctly or not. The mean delay size by presentation are shown in Figure 2.14. Once again, the conditional values reveal that adaptive delays sizes were bimodal: larger for

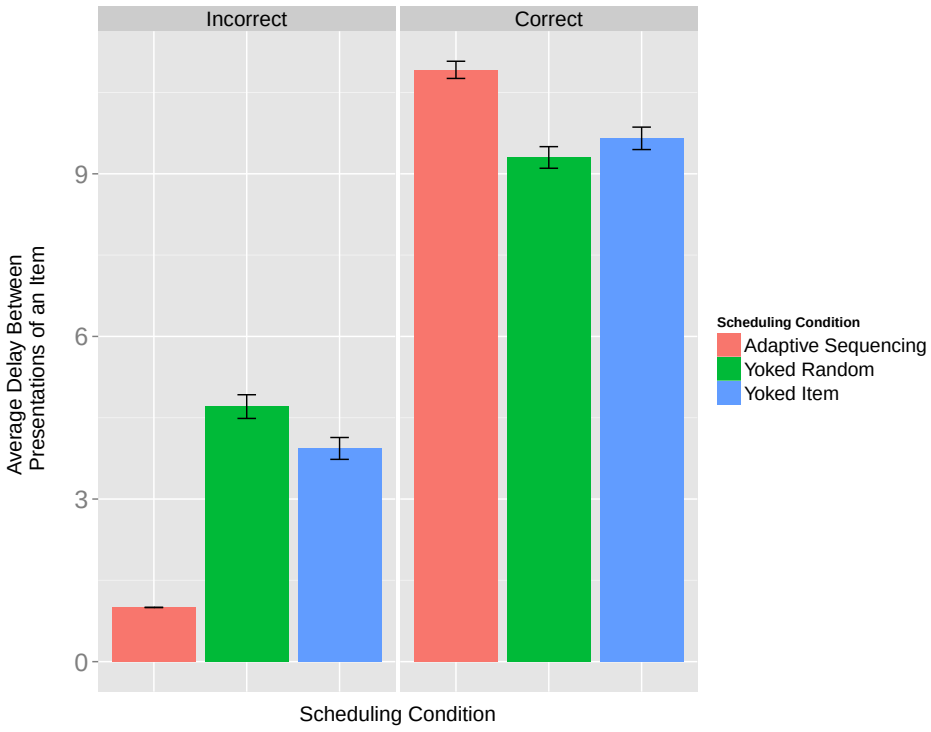

Figure 2.14: Exp 2. Average delay size (in trials) across 3 scheduling conditions conditional on whether the trial preceding the delay was answered correctly or not. Error bars show +/- 1 standard error of the mean.

correct responses (M=11.76, SD=5.04) and smaller for incorrect ones (M=1.0, SD=0.0).

Finally we examined the average delay at each presentation number for the adaptive condition. The mean sizes of the three delays in the adaptive condition are shown in Figure 2.15. The mean initial delay was the smallest (M=1.67, SD=0.98), the second delay largest (M=10.42, SD=2.12), and the third delay smaller than the second delay (M=8.22, SD=1.82). Again it appeared that the pattern of retrievals was not expanding, but expanding-then-contracting. There was only 1 adaptive participant who had strictly expanding practice - positively increasing average delay sizes at each presentation.

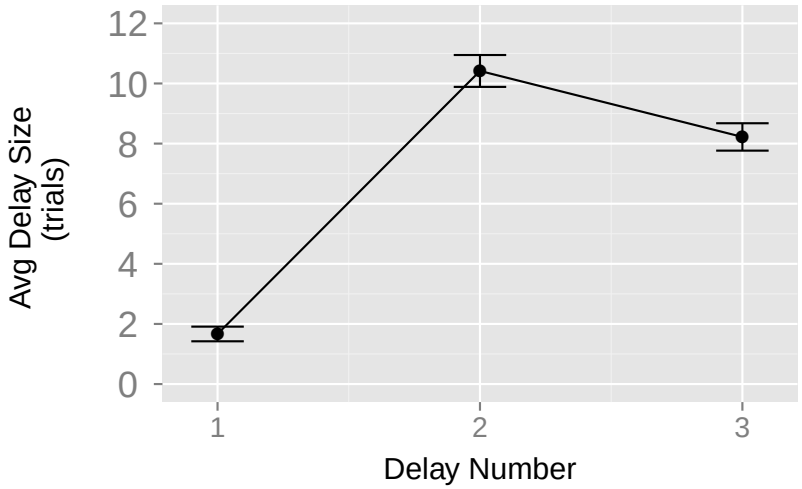

Figure 2.15: Exp 2. Average delay size (in trials) across 3 delays in the adaptive scheduling condition. Error bars show +/- 1 standard error of the mean.

2.3.3 Discussion

In a second experiment comparing adaptive vs. fixed schedules of practice, adaptive scheduling led to larger learning improvements at both an immediate and delayed post-test than yoked schedules. Yoked schedules were unlike adaptive schedules in that 1) delays were not adaptively generated (that is schedule dynamics did not respond to participants' performances), 2) in the yoked-item case delays were not tuned to individual learners, and 3) in the yoked-random case, delays were not tuned to individual items. Adaptive scheduling showed significantly greater learning as measured by change scores between pre-test and post-test than both yoked conditions at an immediate post-test. Adaptive scheduling also significantly outperformed the yoked-random condition, both in terms of greater learning accuracy at both post-tests, and in terms of change-scores and gain scores at a delayed test. In addition, there were numerous marginal effects of adaptive over both yoked-item and yoked-random conditions; gain scores were marginally better at immediate post-test and accuracies were marginally better at delayed post-test for the adaptive condition than the

yoked-item condition, and in all cases, yoked-item performance trended numerically lower than performance in the adaptive condition. In no case, neither at immediate or delayed post-test and not for any measure of performance were the two yoked conditions reliably different from one another.

In summary, participants who learned items using adaptive scheduling had higher performance at immediate and delayed post-tests than participants who saw yoked schedules, where yoked schedules were identical to adaptive schedules in terms of schedule characteristics (average delay size and pattern of delays). Differences between yoked and fixed conditions persisted at a delayed post-test, and yoked schedules were especially poor when they did not preserve spacing delays for individual items (yoked-random condition).

The learning advantages of adaptive conditions over both fixed yoked conditions suggest that the advantages of adaptive scheduling are specific to the ongoing adaptive characteristics of the algorithm. Namely, it is the adaptive algorithm's ability to deal with variability in ongoing learning strengths for learners and items that enables higher long-term learning in the adaptive case, and it is the inability of schedules in the yoked conditions to track and account for changes in ongoing learning strength for learners and items that disable performance in those conditions. The specific advantages of adaptive schedules over the yoked-random condition suggests that adaptive sequencing is driven by identifying the delay schedules most suited to both particular stimulus characteristics and potentially the interaction of individual learners and stimulus characteristics. The discussed findings are noteworthy in that they demonstrate support for some of our stated hypotheses, they speak against some previous hypotheses about spacing effects in the literature, and they open the door to future research on the spacing effect and adaptive schedules.

Evidence supporting hypotheses Greater long-term learning for learners who used an adaptive algorithm demonstrates that the algorithm was able to track and account for changes in ongoing learning strength across learners and items during the learning session. In both yoked conditions, new learners received identical delays to old items and old learners.

However performance in the yoked-item condition declined since learners likely required different scheduling dynamics than old learners. Relatedly, performance in the yoked-random condition declined since individual learning items likely required similar scheduling dynamics to prior learners, but did not receive them. These deficits across the two yoked conditions were regardless of the fact that the absolute delay sizes were identical between both yoked and adaptive participants. Tuning delays to the requirements of learners and items was the goal of an adaptive scheduling system, and these differences between adaptive and yoked conditions show that the adaptive system largely met those goals.

The differences in performance between the adaptive and yoked-random conditions suggests that item characteristics are more important than learner characteristics in the operation of an adaptive scheduling algorithm. Identifying item characteristics, namely predicting intervals of practice for those items, are the heart and soul of adaptive sequencing and it is unsurprising that fixed schedules which do not utilize those characteristics result in poor learning performance. However, the determination of optimal item characteristics alone is probably not sufficient to engender learning. For instance, fixed schedules that have been tuned to item characteristics alone - e.g., in textbooks and other static curricula - would probably not generate greater learning than adaptive schedules of practice for those items. These results also suggest that systems that run prior studies to obtain parameters for their models (Atkinson, 1972; Pavlik & Anderson, 2008), may well have success across different learners. However an important difference is that the ARTS system tested here was able to acquire these parameters while *in use* with learners. Since, in real settings, it would often be impractical to run a prior experiment with similar learners in the same material, there are clear advantages to an adaptive learning system that does not require such prerequisite effort. Finally, the kind of adaptation effects seen here would likely be larger wherever there would be expected to be greater diversity both in learning items as well as in the diversity of learners (our participant pool and our learning materials may have been relatively homogenous with respect to this particular adaptive scheduling advantage).

Evidence against competing hypotheses. This experiment goes a long way towards proving that characteristics of delay size alone, such as average delay size, absolute delay length, etc., are insufficient criteria for a theory of spacing effects. The results of the study make a clear and compelling point: absent a theory of the power of repeated retrievals under particular conditions of delay or difficulty, and absent any underlying notion of ongoing learning strength during learning, rules about spacing effects that focus on delay characteristics applied uniformly to all items are mostly incorrect. For example, if it was the case that delay size during learning was crucial, one should expect the same performance in two conditions that receive the same pattern and size of delays. In fact, we have demonstrated that the opposite occurs: that even when schedule characteristics are equated, learning suffers in the condition where spacing delays do not match item and learner characteristics. The strongest effects involve item characteristics. Additional effects, mostly evident here in immediate posttests, involve differences across learners that can be captured by adaptive algorithms. Similarly, if it was the case that average delay sizes were most important in promoting learning, yoked schedules and adaptive scheduling should have generated similar learning outcomes. Clearly, the primary theoretical reason to expand intervals during practice is to match the characteristics of ongoing learning strength, not to meet particular delay characteristics or criteria of spacing schedules in the abstract.

2.3.4 Exploratory Analyses

2.3.4.1 Increase in failure after spaced delays for non adaptive conditions

One important analysis that we conducted, both during piloting of the study as well as after the study itself, was to determine the relative amount of failure (or 'snapping') that occurs after adaptive delays. Since the theory guiding implementation of the adaptive sequencing algorithm dictates that delay intervals should be just long enough to support difficult retrieval - but not so long as to cause retrieval failure - it was important to assess that, for a range of generated spacing delays, retrieval success stays relatively high. We did this by tabulating

the amount of retrieval successes after adaptive delays (as well as after enforced delays), for each presentation number during the learning session. The proportions of success and failures after adaptive delays are plotted in bins according to size of adaptive delays, creating a histogram, at every presentation number, and for each of the two possible prior response events: whether prior to the delay the item was answered correctly or incorrectly.

These histograms are displayed in Figure 2.16 for the adaptive condition. What is evi-

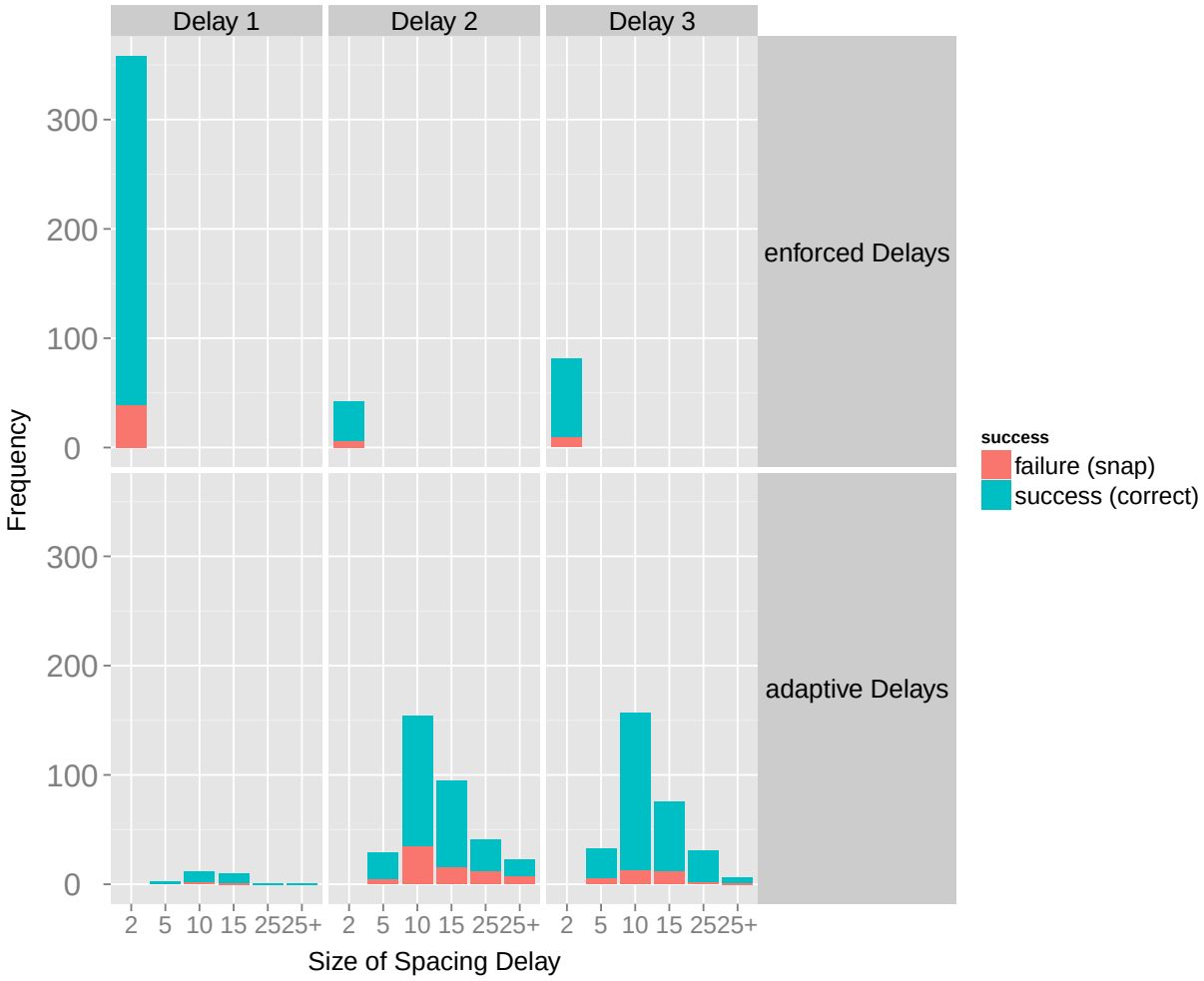

Figure 2.16: Exp 2. Histograms of success or failure (snaps) after adaptive delays by delay number and prior response accuracy in the Adaptive condition.

dent is that, across a range of values of adaptively generated delay, the proportion of response success remains high, whereas the proportion of response failure remains low (adaptive con-

dition: proportion of snaps after incorrect response: 0.11, after correct response, 0.17). This high proportion of successes is evident at even the largest delay sizes, and at all presentations during the learning session. This is exactly the behavior one might expect if adaptive intervals were correctly predicting delays that are long and difficult, but that ultimately result in successful retrieval. The contrary pattern would mostly likely show that, as delays get larger, the proportion of failures, or snaps, gradually increases. This maladaptive pattern is exactly the resulting behavior that we found in pilot testing with incorrect sequencing parameters, where delays were made on average too long for participants. It is also similar to the behavior that is demonstrated in the two yoked conditions. Figures 2.17 and 2.18 show the same response histograms for participants in the two yoked conditions. What is

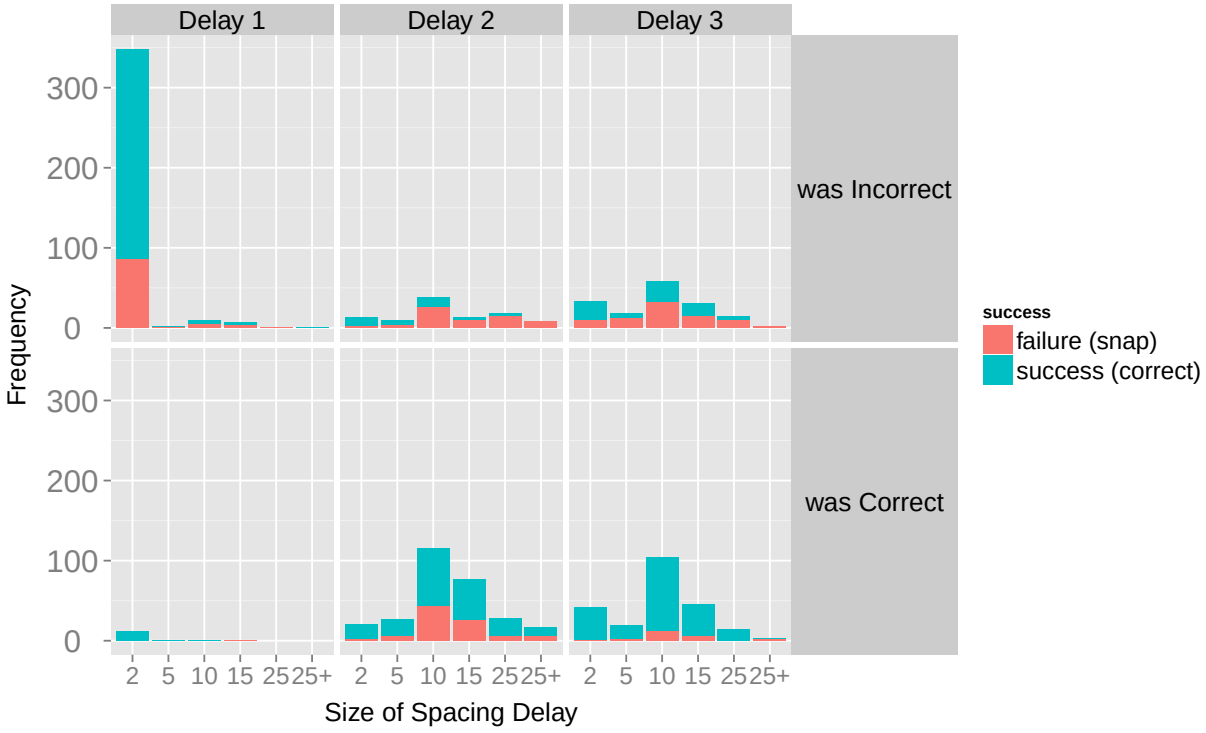

Figure 2.17: Exp 2. Histograms of success or failure (snaps) after yoked delays by delay number and prior response accuracy in the Yoked-random condition.

evident is that response successes and failures do not remain stable across increasing delay sizes. In fact, the proportion of failures is much larger at many delays for participants in the

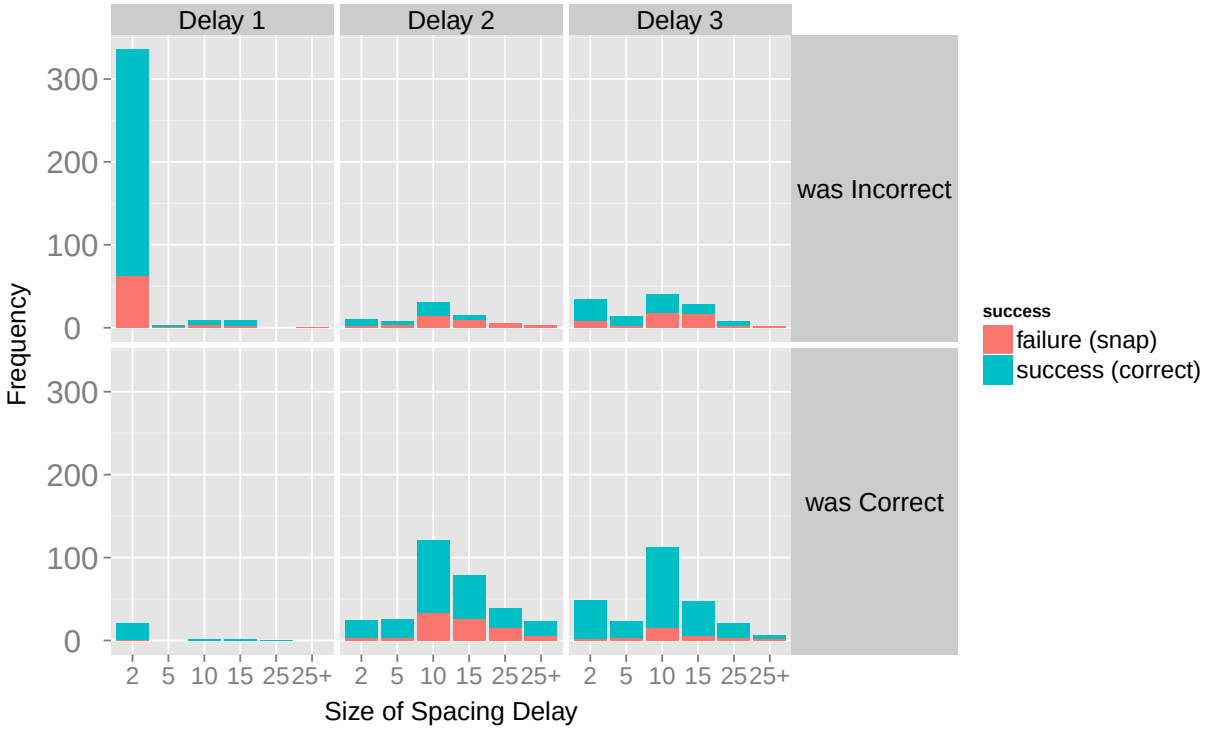

Figure 2.18: Exp 2. Histograms of success or failure (snaps) after yoked delays by delay number and prior response accuracy in the Yoked-item condition.

yoked conditions (yoked-item: proportion of snaps after incorrect response, 0.29; after correct response, 0.20; yoked-random: proportion of snaps after incorrect response: 0.39, after correct response, 0.22). This is consistent with an account of variable learning strength during individual delays - with resulting failures of response after delays that were not designed to support ongoing learning strength for learners and for items.

An inference can be drawn between these results and the poor delayed post-test performance of participants in the yoked groups. We reason that successes at increasingly long delay intervals are crucial to the successful operation of almost any scheduling technique, and that adaptive techniques better enable exactly those successes at exactly those types of delay intervals. Although this observation is only correlational and not directly causal we explore it more in the following additional analyses.

2.3.4.2 Correlation of slope of delay sizes with delayed post-test performance

Finally, we also conducted an analysis that attempted to determine which qualitative schedule of performance – e.g., expanding, contracting – best supported long-term learning. First we computed a measure of 'delay-slope,' the slope of a line of best fit through the three delay interval sizes. Positive delay slopes indicated 'expanding' schedules. Negative delay slopes indicated 'contracting' schedules. Slopes near 0 indicated flat or equal spacing schedules. Of course, the adaptive algorithm, as noted in prior analyses, rarely generated strictly equal or expanding schedules (more often than not schedules were expanding-then-contracting), but we felt that constructing a continuous dimension to measure the relative pattern of delay sizes was a highly valid statistical abstraction and would reflect well on prior experimental efforts to divine the benefits of qualitatively differing schedule types. Correlations between delay slope and delayed post-test performance are illustrated graphically in Figure 2.19 for each condition. Data points were for individual items: predictors were the delay slope of items and the outcomes were binary delayed post-test accuracies (accurate or inaccurate). To simplify the analysis, we chose linear rather than logistic regression. Correlations between delay slope and delayed post-test performance were positive and significant for the adaptive condition (slope b=0.036, t(382)=3.91, p<.001; R=0.20), and for the yoked-item condition (slope b=0.017, t(382)=1.98, p=.049; R=0.10), but not significant for the yoked-random condition (slope b=0.003, t(382)=0.34, p=.74). Positive correlations imply that increasing slopes – that is, more highly expanding schedules – resulted in greater post-test performances. This was most evident for the adaptive condition, slightly less so in the yoked-item condition, and there was a flat pattern and no significant correlation for the yoked-random condition. This result implies first that in the adaptive condition, schedules that tended to expand the size of delays across presentations were correlated with better long-term learning. It also shows that although some schedules tended to expand in both the yoked-item and yoked-random conditions, expansion enough was not enough to correlate with better delayed post-test performance in the yoked-random condition. In the yoked-item condition, the correlation was significant but the slope of the correlation was less than in the adaptive

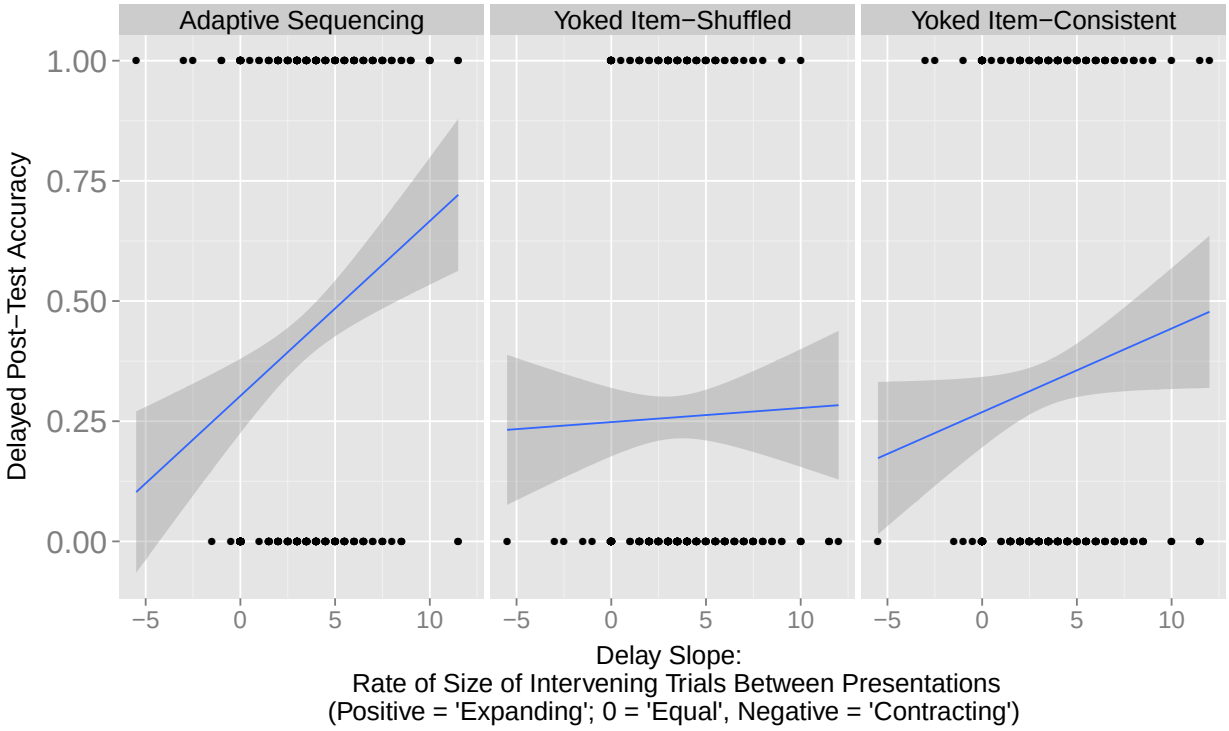

Figure 2.19: Exp 2. Delay slope correlated with delayed post-test performance, across 3 conditions. Data points were accuracies on delayed post-test and delay slope of individual items. Delay slope was computed using a line of best fit through an item's three delay values. Delay slope was positive if an item had expanding delay intervals, negative for contracting intervals, and 0 for equal intervals. Shaded regions show a 95% confidence interval.

case. It should be noted that the R squared values were fairly low. The takeaway from this analysis is that expanding schedules can be correlated with better learning gains, but mostly in cases where spacing delays respect ongoing learning strength, e.g., in an adaptive scheduling condition. This result weighs in partially on the debate between qualitative schedules of practice, and it shows that even among schedules that differ only slightly in the qualitative schedule of practice, there are differences in whether particular qualitative schedules generate learning advantages. Thus, the match between ongoing learning strength and the benefits of a particular pattern of spacing delays participate cooperatively in generating greater learning.

2.4 Conclusion - Experiments 1 & 2

Across two experiments, we showed that a technique of adaptive scheduling in order to maximize the spacing delay between presentations of items was successful in producing greater learning benefits than related fixed schedules of practice. In both experiments items were presented for the same limited number of presentations (four), where the only manipulation was to the order of presentation of items during the learning session. Experiment 1 demonstrated that learning gains were higher for an adaptive scheduling condition than a series of fixed schedules commonly used in the research literature of studies on learning and memory. Experiment 2 demonstrated that even when schedules were equated in the size of the spacing delays across conditions, adaptive schedules outperformed fixed ones. Specifically the locus of the improvements appeared to be over those schedules whose spacing delays did not meet item and learner characteristics. In both experiments the improvements were measured at both immediate and delayed post-tests. It appears that the strongest benefit of adaptive scheduling is to track the ongoing learning strength of variable learners and items, and that the resulting schedules are more powerful at promoting learning than most fixed schedules of practice.

It is important to step back and qualify the contributions of the present series of experimental schedule manipulations to larger concepts and theories in the study of memory and learning. Although a purpose of the preceding experiments has been to study the benefits and limitations of particular scheduling strategies among a set of competing ones, a further purpose has been to determine the locus of learning benefits, and the nature of learning mechanisms more generally. These general findings in some cases take the form of new dimensions previously not addressed in theory (for example, the degree to which adaptation affects learning). In other cases, the processes are directly relevant to pre-established theories of learning and enable us to evaluate and assess those theories.

Other issues - Total presentations It should be noted that adaptive sequencing was originally designed to support learning to criterion - that is, the continuation of presentation

of any item until the point that a learner meets a learning criterion for that item, also known as 'drop-out', 'retirement', or 'criterion-based learning' (see Chapter 3). Instead, in this study, all conditions were designed so that each item received four total presentations in the learning phase. This is a highly artificial way of producing real learning - likely not to be used in practice by any actual student or learner. No learner would be expected to be satisfied with the fairly low resulting learning performance that resulted at delayed test, and no scheduling condition given only 4 total presentations per item would be able to appreciably increase those low learning rates, even if a mathematically optimal learning schedule was found. There are simply not enough presentations to support robust, long-term learning. To researchers credit, it is probably assumed that such controlled studies can inform us about theoretical aspects of the spacing effect, in a way that realistic studies cannot. It is assumed that controlled conditions such as limited presentations (often presentations are limited to only 2 tests with a single spacing delay), can provide quantitative information that informs the design of more realistic, longer-lasting teaching and study strategies. Unfortunately, there is less evidence of productive transfer of knowledge from controlled studies to real-world learning situations than would be desirable. It is hoped that in future studies, this gap can be bridged. By conducting work on learning schedules of a more realistic nature, and uniting work from highly constrained studies with more realistic studies, better hypotheses about the conditions of practice can be formulated in both domains.

On the whole, we think these experiments bring research on the spacing effect steps closer toward the goal of depicting the gamut of spacing effects through the relative strengths and weaknesses of particular schedules of practice, and through sophisticated measures of performance, a strategy outlined by Landauer:

> "...using information theory to model memory functions [is] a powerful methodological tool I believe to have been woefully underused in cognitive psychology (despite the loosely labeled 'information processing' approach). For spacing effects in particular, an example of what might be done is to study the cumulative storage efficiency of differing repetition regimes and scheduling distributions.

> This might yield more accurate, quantitative, and continuous valued treatments of the effects of variables and their interaction than measures of raw frequencies of right and wrong answers or raw accuracy of performance." - (Landauer, 2010)

Of course, we are not here using information theoretic rules to describe learning gains in these experiments, but we are conscious of the goal of contrasting the likelihood of current, possible learning gains with the space of possible ones - a strategy not unlike the tenets of information theory - and we are of course seeking dimensions of influence on spacing effects, including effects of adaptation.

What remains to be done is to test more realistic learning schedules, of longer duration and with stopping rules that entail learners reach reasonable levels of proficiency with material, and to assess their encoding efficiencies - that is, to examine how quickly learners reach learning criteria - to determine which schedules require the fewest number of presentations to achieve durable learning. We turn to more realistic learning situations in Chapter 3

CHAPTER 3

Adaptive vs. Random Scheduling

Comparing Schedules using Realistic Learning Conditions

3.1 Introduction

3.1.1 Limitations of Fixed Schedule Studies

In Chapter 2 we demonstrated that the technique of adaptively expanding delays between trials during a learning session is a beneficial strategy for increasing the short and long-term retention of basic facts. These studies do not, however, definitively prove the effectiveness of adaptive schedules in typical real-world learning settings, where learning sessions are not limited to a few presentations per item, and where it is usually desired that presentation continues until learning has reached some learning criterion or standard of proficiency (Bloom, 1968; 1974).

It is therefore important to demonstrate that the advantages of adaptive schedules are general; that adaptive schedules also operate effectively under realistic learning criteria and do well compared to common learning strategies. What are some reasons that adaptive schedules would not perform as well under general conditions of learning? One possibility is that learning criteria may make up for the deficiencies of a non-adaptive schedule, counteracting the poorness of the match between spacing interval sizes and ongoing learning strength. Alternatively, matching spacing delays to ongoing learning strength may simply

be less important or less effective at later points in the learning session; adaptive sequencing may simply be less effective at generating the last, longest delays for an item before it reaches a learning criterion. Finally, adaptive schedules may not compete well with some commonly used schedules - such as random presentation schedules - that incorporate a variety of robust spacing delays as well as learning criteria.

To address these questions, in this chapter we set out to investigate adaptive schedules of practice using learning to objective criteria, and to compare adaptive to random schedules under more typical durations and conditions of learning practice. First we assess some background literature on the topic of learning criteria, dropout and efficient learning, and we assess their relationship to spaced schedules of practice. We then present two new studies that address the relative benefits of adaptive vs. random scheduling, both with and without learning criteria.

Learning efficiency In order to investigate these questions we must more closely consider the character of the learning performance that is affected by realistic durations of practice, and features such as learning criteria and dropout. Crucial to our investigation will be a performance measure of 'efficiency', that is, the number of learning trials invested in a learning session in relation to overall performance (accuracy) gains on a post-test. Only with respect to trials invested will learning accuracies be interpretable, since we can predict that different schedules (e.g., schedules with differing learning criteria across conditions) will deliver differential amounts of practice on average across conditions. We consider dropout and then learning criteria and spacing in light of a measure of efficiency.

3.1.2 Learning Criteria, Dropout, Spacing and Efficiency

3.1.2.1 Overview of learning to criteria

What are learning criteria and what relationship does a learning criterion have with dropout? By learning criteria it is generally meant that some number of successful recalls of an item

are necessary in order to demonstrate proficiency with that item. Although seldom tested, learning criteria can also incorporate fluency criteria, e.g., responding that meets preset speed criteria. By 'dropout' it is meant that an item is no longer presented for the duration of the learning session. Dropping out items from the learning set encourages 'efficient' learning: practicing items fewer times but with greater benefit to learning. An obvious dropout strategy is to remove an item from the learning set at the moment it reaches a learning criterion.

Generally, learning criteria are related to the notion of mastery learning (Bloom, 1968; 1974), or the notion that individuals should master units of learning before proceeding to later learning. Mastery learning treats learners as individuals, having different requirements and taking differing amounts of time to achieve unified learning standards. This is as opposed to much traditional instruction that grades students on units of instruction rather than ensuring successful learning. Mastery learning was the motivator of early attempts to create adaptive learning techniques, for example in curricula such as 'programmed instruction' (Skinner 1954; Holland and Skinner, 1961). As principles of cognitive science get applied to the real world, mastery learning has become considerably more important. In addition, technological advances have made tracking of individual learning items and assessment of instructional objectives easier to achieve. Any future, realistic application of spacing to learning should be compatible with, and studied in the context of, learning to criterion.

How do learning criteria affect learning? Criterion learning has been studied generally in terms of its effects on learning, as well as specifically in terms of the kind or amount of recall success required to meet a criterion.

In studies of memory and learning, criterion learning appears to improve all aspects of associative retrieval (Vaughn & Rawson, 2011). Specifically, as Rawson and Dunlosky have reported (2011), increases in the stringency of learning criteria have a logarithmic relationship with later recall; that is, stronger criteria result in greater amounts of learning (as measured by retention on a delayed post-test), but there are diminishing returns on learning for every unit increase in the strictness of a criterion. In other words, there are

limits to how strict a learning criterion should be. Rawson and Dunlosky (2011) manipulated the number of successful recalls required before items dropped out of a learning session. They repeatedly tested participants with passages of textbook material and learners 'self-monitored' the success of recall of key facts. Items were considered learned after either 1, 2, 3, or 4 successful recalls, after which items were no longer shown to participants. Rawson and Dunlosky found that greater criterion level (2, 3, or 4 recalls) increased the success of recall on a test after 2 days, but that the difference between an initial criterion of 3 vs. 4 recalls was not significant. It also appeared that given a second re-learning session where items were learned to the same criterion level, the differences in recall between all criteria levels vanished in a third, delayed session. Finally, their analyses showed that the largest difference in delayed recall given initial criterion level came from extremely poor learning in the one-correct response criterion case, suggesting that a single correct recall is either a poor criterion, or perhaps categorically different from a true learning criterion. Thus learning criteria do appear to affect learning in a positive but negatively accelerating way. However the conditions under which criteria affect learning are highly constrained. It is also possible that, if measured purely in terms of the number of successful recalls, the presence or absence, rather than the exact level or strictness, of a learning criterion has the single greatest effect on learning.

Compared to learning gains seen without a learning criteria in fixed comparison studies, would adaptive schedules convey similar or greater learning gains when sessions extend beyond a few short presentations? One possibility is that they will not. For example, after the first few presentations, it may be that the impact of adaptive delays on learning differs from the impact during the first few presentations. Adaptive delays may not help much during later learning, or the power of an adaptive algorithm to accurately predict delay size based on response speed may diminish. Relatedly, the advantage of adaptive delays over other types of schedule may diminish when there are enough encoding opportunities to achieve reasonable learning criteria, regardless of the match between spacing interval sizes and ongoing learning strength. In other words, learning to a criterion may make up for the deficiencies of

a non-adaptive schedule. Alternatively - and as we predict - an adaptive schedule of delays may in fact produce even greater learning or - crucially - even faster, more efficient learning, that is, using fewer total trials presented per item. The reasons for this advantage fall largely in the scope of the theoretical background discussed in Chapter 1: increases in difficulty at each recall produce even greater gains in long-term storage strength, enabling spacing delays to expand as learning progresses. Accurately predicting the length of spacing delays ensures continuously difficult, but successful retrievals across the learning session. More numerous successful retrievals and fewer failed retrievals add up to fewer total learning trials and thus greater learning efficiency.

3.1.2.2 Learning Criteria and Spacing

One point not fully addressed in the prior literature is how learning criteria might interact with spacing between learning events. To our best knowledge there is only one existing study that has systematically manipulated both learning criterion and spacing of learning events to examine the interaction of spacing and criterion. Pyc and Rawson (2009) had participants learn swahili-english word pairs under two spacing conditions (long and short spacing) and a number of criterion levels - roughly 1 to 10 correct retrievals. Retention intervals were also manipulated. Long spacing intervals generally improved learning, but increases in criterion level showed curvilinear decrease of learning gains at both spacing durations. In other words the difference in gain between learning criteria of either 9 or 10 correct recalls was not as great as between 2 and 3. Their finding is roughly in line with that of Rawson and Dunlosky (2011), that increases in criterion strength result in logarithmically decreasing improvements to final recall performance. In relation to the issues of spacing and criterion that we have been discussing, the limitation of Pyc and Rawson's (2009) study is that the effects were only investigated under the scheduling constraints of equal spacing, rather than expanding or other forms of spacing. Consistent with a theory of retrieval difficulty that supports expanding retrieval practice, equal spacing schedules may not provide consistent levels of difficulty at each practice, thus reducing the difficulty of nominally more 'stringent' criteria levels. As

an alternative, increasing the spacing across criterion trials might counteract the negative acceleration of learning strength by maintaining high levels of difficulty at every criterial learning practice [1]. With increases in spacing accompanying increases in the difficulty of a criterion, there may be super-additive effects as opposed to diminishing returns: the form of the relationship between criterion difficulty and long-term learning may be linear or even exponential.

Indeed, it will be important to assess the difficulty of the ensuing retrievals during criterial practice. For example, a continuous sequence of massed repetitions of an item likely promotes perfect performance in the immediate term, no matter the strictness of the criterion. Similarly, a schedule that promotes extremely difficult material may find that criterial performance is nearly impossible to achieve. Thus criteria need necessarily be evaluated in relation to the kinds of ongoing schedule dynamics during the criterion trials, namely the spacing of learning events.

In the adaptive algorithms we have introduced in this thesis, we have devised learning criteria that implicitly contain spacing criteria as well as explicit speed and accuracy criteria. Since our learning criteria require fast responses in addition to correct responses (responses must be made in less than 7 seconds), and since adaptive spacing intervals are a function of the speed of response, the resulting criteria enforce spaced, difficult retrievals, not just correct or fast ones. One of the goals of the present research was to examine the benefits of such learning criteria on learning.

As noted, learning criteria that incorporate response speeds are a natural feature of adaptive scheduling algorithms of the sort tested here. The incorporation of response speeds into learning criteria is exceptionally rare in practice and in the experimental literature. Using response speeds in learning criteria is one the benefits of using computer based learning

[1] Possibly one of the reasons experiments along these lines have never been conducted owes to the general lack of knowledge about the exact mathematical relationships between spacing and criteria. Doing such an experiment may be particularly risky if the match between item spacing and ongoing learning strength is weak - that is, the effect of a strict criterion may be lost if spacing does not support difficult retrieval. Although we do not attempt a systematic exploration of the interactions of spacing and criterion level in the current studies, we assume that adaptive spacing enables a wider range of learning criteria to convey high levels of benefit to learning.

technology to schedule learning. Since much of the theory behind the operation of the adaptive sequencing algorithm places response speed in a central position in the determination of learning strength, it is natural that response speeds should also form part of learning criteria: knowing how quickly responses are made can be as informative as knowing how sufficiently material has been learned. The use of fluency criteria goes beyond the mere value of time saved by fast responding. In many complex learning domains, basic tasks must be relatively automatic in order to allow higher level skills to be built on top of them. RT criteria in learning provide some indication of automatic processing (Schneider & Shiffrin, 1973). Just as with accuracy learning criteria however, response speeds should be chosen within the context of ongoing retrieval difficulty and spacing delays. Response speed-based learning criteria will only be effective if the conditions of practice support them. There is a bright future for an experimental analysis of response speeds in learning criteria and we broach this topic briefly in Chapter 4, experiment 3.

3.1.2.3 Learning Efficiency

Use of learning criteria introduces a problem in comparing the effectiveness of learning conditions. Because individuals may reach the set criteria after different amounts of time or numbers of learning trials, learning effects involve two dependent measures, rather than one. Accuracy as seen in a post-test is one, and the number of trials to attain a given accuracy level is another. Thus, learners in one group may take longer to learn and perform better in terms of post-test accuracy than another group that requires fewer learning trials and performs less well. Which is better? How can we compare learning conditions when two dependent measures are involved in order to obtain "apples to apples" comparisons? Put another way, in most real-world settings, both variables – how long learning takes and what gains it yields – are important. In fact, some measure of learning rate – how much is learned per unit of investment – might be most important. In the present work, we use two methods to obtain "apples to apples" comparisons. One, which we describe in more detail later, is to examine learning at or close to the point when participants in one learning condition (the one

requiring fewest trials) are at the point or close to the point of reaching the learning criteria. We can look at performance in all groups after that number of trials. (Section below.) This measure is helpful, but it has some limitations, including that learning to criterion has not yet occurred for some participants when it is applied, and that it is not applicable to delayed tests, which may be more important for gauging lasting learning.

Here we describe the primary method for comparing conditions that may vary both in learning gains and in the number of learning trials required to achieve them: We use a measure of learning efficiency that combines these dependent measures. Crucial to our investigation will be a performance measure of 'efficiency', that is, the number of learning trials invested in a learning session in relation to overall performance (accuracy) gains on a post-test. Efficiency is shown in equation 3.1, where A_p is the accuracy on a post-test and T is the number of learning trials invested.

$$e = A_p/T \tag{3.1}$$

Only with respect to trials invested will learning accuracies be interpretable, since we can predict that different schedules (e.g., schedules with differing learning criteria across conditions) will deliver differential amounts of practice on average across conditions. We consider dropout, learning criteria and spacing in light of a measure of efficiency.

3.1.2.4 Dropout and Learning Efficiency

How does 'dropout', that is, removal of well-learned items during the learning session, affect learning performance? Kornell and Bjork (2007) found that dropout was an intuitive, common strategy that most students employed while studying, but that when used in practice, students' inability to accurately monitor their own learning strength lowered their overall learning gains (Kornell & Bjork, 2008). In other words, students' recall performances were poorer when they chose to drop learning items. When dropping items based on learning criteria, Pyc and Rawson (2011) showed that learning improved - as measured by efficiency - compared to fixed schedules of practice where there was no dropout. In terms of raw recall

accuracies, performance was generally better for fixed schedules than for dropout schedules; however two points should be kept in mind. First, Pyc and Rawson's criterion for dropout consisted of one correct response. Second, a conditional analysis showed that the primary advantage for fixed schedules was when performance for items was initially correct earlier in the schedule, suggesting that many items in fixed schedules experienced learning beyond the criterion and that many items in dropout schedules were dropped too early. These studies and some intuition indicate that presumably, the gains in efficiency found from dropping items could be balanced with the losses incurred from when items are dropped too early. Again, one of the purposes of the studies in this chapter was to examine schedules of practice under realistic learning conditions - using features that make sense in terms of long-term learning, e.g., strict criteria (five correct responses in a row) and criteria that ensure recall is difficult, by spacing across long time scales.

We think there is an important definitional point to be made about the use and study of dropout in learning schedules. As mentioned, dropout schedules are one side of a coin that includes learning criteria (and in some sense, retrieval difficulty). Dropout cannot happen except in light of a measure of learning strength, whether through a students' poor estimate via self-monitoring, or through criteria or rules that are assumed to monitor the development of learning strength. Criteria that poorly estimate learning strength are not likely to show an effect of dropout - or conversely may show negative effects. It may be the case that the true question of when and how dropout should be used should be focused on the development and understanding of proper learning criteria rather than the development of techniques or rules for enacting dropout.

In addition to learning criteria, dropout may be intimately related to learning efficiency. Dropout may make learning more efficient by ensuring that some items cease practice earlier than they would have given unmodified presentation. However, because of this relationship, it may be easy to erroneously conclude that drop-out always improves efficiency. A number of factors make it less than certain that long-term efficiency will improve with dropout. First, there are benefits to practice beyond meeting learning criteria. This effect of improvements

to learning beyond nominal learning criteria is known as over-learning (Underwood, 1964; Krueger, 1930). Because learning can still improve beyond the point that a given performance criteria is met, it is not guaranteed that dropping items after meeting criteria maximizes long-term efficiency. Further, since learning criteria provide somewhat different levels of difficulty of retrieval practice for items, items may prematurely drop out if all items share learning criteria (Vaughn, Rawson, & Pyc, 2013). In addition, dropout is likely to affect efficiency differently for schedules of varying difficulty. Overall it is important to note that dropout and efficiency are directly tied, but it is not a given how efficiency will be affected by changes in the way dropout is conducted.

Monitoring ongoing practice to determine if learning criteria have been met is a good example of using information about the learner's performance and progress to arrange what happens in learning - in other words, to adapt learning. It is worth stressing that learning criteria are part and parcel of an advanced, adaptive way to schedule learning, that is, learning criteria (and dropout) are adaptive learning features. Since it will be necessary to design learning systems around these features in the future, it is important to study them, and develop strategies that best implement learning criteria for a wide range of learning situations.

3.2 Motivation for Studies and Key Questions

In summary, spacing effects in cases involving learning criteria have not been extensively studied, and the effects of dropout have not been tested much in connection to adaptive learning techniques. In addition, we know of no previous studies that have used response-time criteria in dropout, much less have used learning criteria in relation to adaptive spacing. What we have done in the following studies is attempt to draw contrasts between obvious and reasonable combinations of such features.

We conducted a comparison between two types of schedule: adaptive schedules and random schedules of presentation. In the adaptive schedule condition, items were presented

using the adaptive sequencing algorithm presented in Chapter 1; however, unlike experiments in Chapter 2, learning continued until each item met a learning criterion, after which individual items were dropped from the set of learning items. In the random schedule case, learning items were presently randomly with no restriction on repetition or length of delay between presentation, and items were not dropped from the session. Each item was tracked so that after every item had met learning criteria the session was terminated. This comparison stood to demonstrate whether learning in an adaptive condition could compete with unmodified random presentation, where neither spacing nor dropout was controlled by the computer. In a second experiment we compared adaptive presentation with random presentation, where the adaptive and random conditions both had identical learning criteria and both schedules included dropout. In both experiments, the learning criteria for every item was five of the last five presentations correct with response-times less than seven seconds. Crucially we looked at the efficiency of learning across the two scheduling conditions, at both immediate and delayed tests of retention.

We aimed to answer the following questions: Do adaptive schedules result in greater learning efficiency than random schedules? Do adaptive schedules result in greater learning efficiency when compared to schedules that have similar learning criteria — that is, does an adaptive schedule show efficiency gains beyond that which is achieved through the effect of a learning criterion? Finally, we assessed whether schedules differed at equivalent points (identical number of trials) during learning.

Random schedules Despite the focus in the experimental literature on organized and theory-driven methods of retrieval practice (e.g., expanding retrieval practice), it bears mentioning why we have chosen random schedules rather than other fixed schedules. Random schedules naturally implement a form of spaced practice, where spacing intervals for items are on average as long as the number of items in the learning set. In addition, random practice tends to encourage highly variable contexts of encoding — that is, each practice with an item is usually preceded by and followed by a different set of items; conditions that,

according to some theories of the benefit of spacing, are hypothesized to have a strong effect on learning (Young & Belleza, 1982; Howard & Kahana, 1999). Probably the connection between conditions of encoding variability and retrieval practice is strong: retrievals would be likely to be more difficult given differing encoding contexts. Interestingly, almost no studies in the literature on scheduling and spacing include random practice as a comparison (in contrast to many studies in the motor learning literature that compare blocked vs. random schedules, e.g., Shea & Morgan, 1979). In addition, truly random schedules do involve some degree of learning "technology" in the sense that purely random schedules are difficult to generate without a computer. In some sense, random schedules are understudied and they may provide a window into extra-theoretical conditions of practice that are beneficial for learning.

There are also reasons for choosing random schedules over other fixed schedules. For instance, with qualitative schedules of spaced practice such as expanding practice, it is not clear how such schedules should proceed. That is, if there is not cessation of a session until learning criteria are met, should expanding spacing continue? How should filler items be introduced in those cases to support longer retention intervals? How could one ensure the proper interleaving of expanding schedules if items are continuously dropping out - altering the schedule of future presentations? For these and other related reasons, random schedules are particularly suitable to investigations that involve learning criteria, drop out and measures of schedule efficiency.

Summary of studies To summarize, we asked whether the benefits that adaptive practice convey to learning extend beyond the simple, constrained learning conditions tested in Chapter 2. That is, we aimed to show that in conditions of practice — namely an extended learning session where items are continuously presented until they meet learning criteria — that adaptive schedules also out-perform typical realistic schedules of practice, namely random presentation. Further we aimed to show that this benefit exists whether or not dropout is considered as a feature of the adaptive system.

3.3 Experiment 1: Random vs. Adaptive Presentation (Random Without Dropout)

3.3.1 Method

Participants Participants were 48 UCLA undergraduates who participated for course credit.

Experiment design The experiment used a pre-test/post-test design, with a delayed post-test administered after 1 week. Pre-test and post-tests consisted of the same 24 questions. There were 2 between subjects conditions, adaptive and random, that manipulated the scheduling of items during the learning session. Items were either presented using an adaptive schedule identical to the adaptive algorithm in Experiment 1 and 2 in Chapter 2, or a random presentation schedule.

Materials Materials were nearly identical to those in Experiments 1 and 2 in Chapter 2. The learning items were the same 24 countries used as learning items in Experiments 1 and 2, but there were no filler items used.

Procedure Participants went through the same procedure as in Experiments 1 and 2 (Ch. 2).

Unlike Experiment 1 and 2 (Ch. 2), in the adaptive condition, there was not a fixed number of presentations of each item. Each item was learned until it reached a learning criterion and was dropped out of the set. In the random condition, each item was tracked so that the experiment session ended after every item had reached a learning criterion, or after the learning session reached 45 minutes, whichever came first. There was no dropout of items during the learning session in the random condition.

The learning criterion was five out of the last five presentations of an item correct with all five response-times less than seven seconds.

The parameters of the adaptive algorithm did not change from Experiments 1 and 2 (Ch. 2).

Dependent measures The primary dependent measure was learning efficiency. Efficiency is a measure of post-test performance that takes into account the total number of presentations a learner receives during a learning session. Because different amounts of exposure to practice items during learning can lead to different amounts of overall retention on a post-test, and because the learning schedules that we used took differing number of trials for participants to complete the learning session, we relied on a measure of efficiency to describe learning performance. Efficiency can be defined as the number of items correct on a post-test divided by the total number of trials in training. In the current experiments, we subtracted the number of pre-test items answered correctly from the number answered correctly on a post-test before dividing by the number of trials in the learning session. Efficiency is thus a rate that expresses the amount (fraction of an item) gained on a post-test for every trial invested in learning. Owing to the proportional nature of the measure, changes which appear small in units of efficiency, for example the difference between .06 and .08, are actually quite large; a difference of 0.02 is actually a 30 percent difference in the example.

Efficiency was calculated based on the change score (post-test accuracy minus pre-test accuracy) between pre and post-tests.

3.3.2 Results

Primary dependent measure: Learning efficiency Results for efficiency at immediate and delayed post-tests are shown in Figure 3.1. A 2x2 mixed factor ANOVA on scheduling condition (adaptive vs. random) and post-test phase (immediate vs. delayed) ANOVA found significant main effects of condition ($F(1,46)=33.8$, $p<.001$), a main effect of post-test phase ($F(1,46)=89.7$, $p<.001$), and a significant scheduling condition by test-phase interaction ($F(1,46)=36.6$, $p<.001$). At immediate post-test, efficiencies were higher in the adaptive condition ($M=0.109$, $SD=0.03$) than the random condition ($M=0.054$, $SD=0.012$)

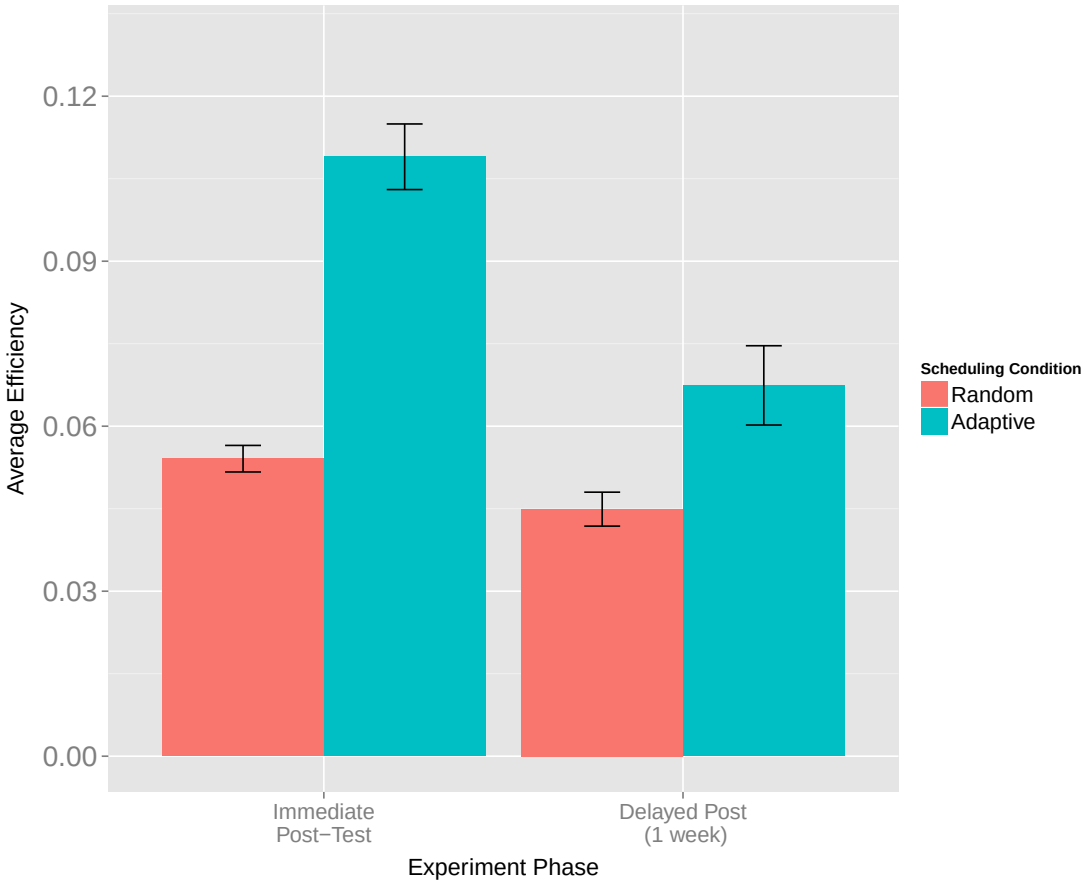

Figure 3.1: Exp 1. Efficiency at immediate and delayed post-tests by scheduling condition. Efficiency was the change in accuracy from pre-test to post-test divided by the number of trials in training. Error bars show +/- 1 standard error of the mean.

a significant difference (t(46)=8.53, p<.001, Cohen's d=2.68). This outcome represents a 102% greater efficiency in the adaptive condition at immediate posttest. At delayed post-test, efficiencies were also higher in the adaptive condition (M=0.067, SD=0.04) than the random condition (M=0.045, SD=0.015), a significant difference (t(46)=2.87, p=.006, Cohen's d=0.892). These differences comprise a 50% greater efficiency in the adaptive condition at delayed posttest. Comparing means between the two test phases, the difference between efficiencies at each test phase for both the adaptive and random condition were significant (adaptive immediate vs. delayed, t(23)=5.88, p<.001; random immediate vs. delayed, t(23)=8.11, p<.001). The interaction appeared to be the result of declining efficiency in the

adaptive condition from immediate to delayed post-test, but smaller decline in the random condition.

Trials to criterion Because learning took differing amounts of time in the two scheduling conditions owing to differing drop-out schedules, the number of trials to criterion was analyzed. The number of trials to criterion in each condition are shown in Figure 3.2. The

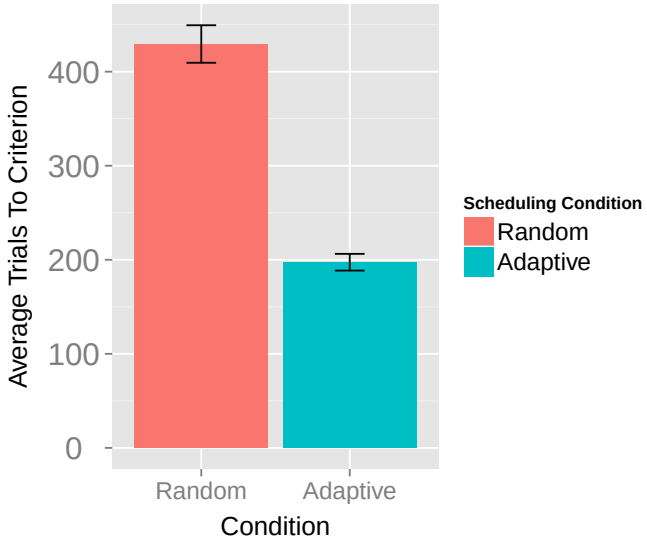

Figure 3.2: Exp 1. Number of trials in learning session for each scheduling condition. Error bars show +/- 1 standard error of the mean.

random condition took on average 429 trials to reach the end of the session (SD=97.7). The adaptive condition took on average 197 trials (SD=43.5). This difference was significant ($t(46)=10.6$, $p<.001$, $d=3.29$).

Equivalent learning trials analysis Because drop-out strategy and schedule structure led to differing numbers of trials during training, we analyzed the two scheduling conditions at equivalent points near the end of training - at the moment that participants in the adaptive condition retired - in order to assess whether learning was still progressing in the

random condition. Comparing at equivalent moments in learning allows for an alternative comparison of training performance than average accuracy at a post-test, where, owing to learning criteria, performance is expected to be the same across conditions.

Accuracies were compared at the moment participants in the adaptive condition finished the learning phase, trial 197. The accuracy of the prior 2 presentations for each item was measured in both conditions. Accuracies are shown in Figure 3.3. Average accuracy was

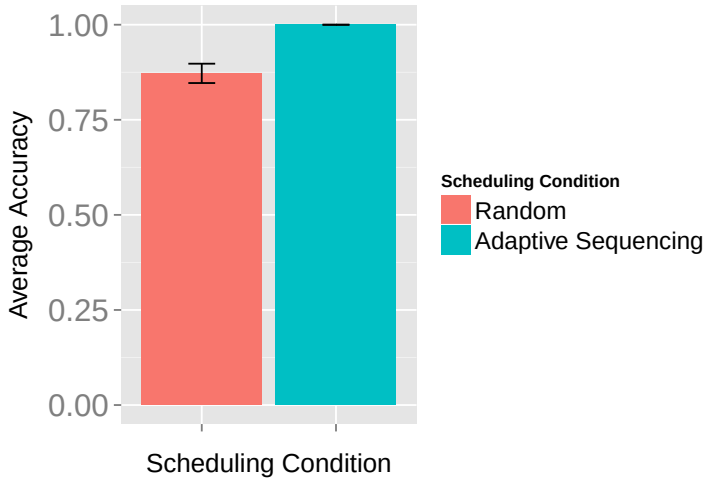

Figure 3.3: Exp 1. Accuracy at equivalent points in learning (trial 197) in both scheduling conditions. Error bars show +/- 1 standard error of the mean.

higher in the adaptive condition than the random condition (adaptive M=1.0, SD=0; random M=0.87, SD=0.12), a significant difference ($t(46)=5.03$, $p<.001$, $d=2.05$).

Average reaction times were also taken at the equivalent point in learning. Reaction times are shown in Figure 3.4. Reaction times were lower for the adaptive condition (M=2.66 sec, SD=0.31) than the random condition (M=3.03 sec, SD=0.65), a significant difference ($t(46)=2.55$, $p=.014$, $d=0.78$).

We also analyzed accuracies at a number of trial blocks before the equivalent learning point. In blocks consisting of 3 trials per item, and starting 5 blocks before the equivalent learning point, accuracies across condition are displayed in Figure 3.5 (blocks are displayed in reverse in the figure). A 2X4 mixed factor ANOVA on schedule condition and trial block

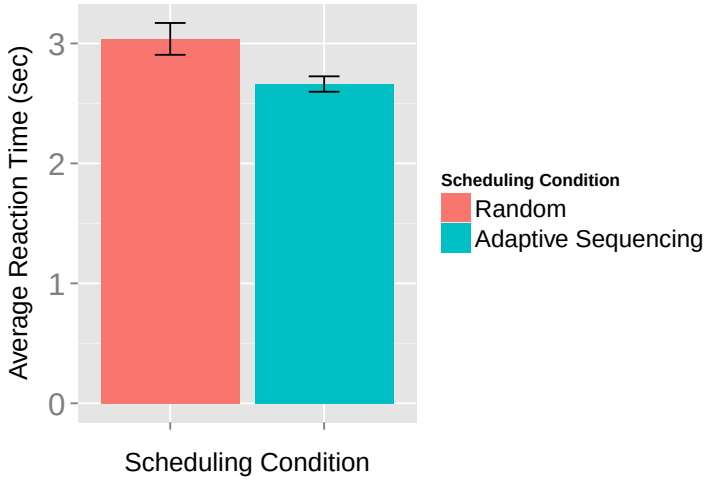

Figure 3.4: Exp 1. Reaction time at equivalent points in learning (trial 197) in both scheduling conditions. RTs are from correct responses only. Error bars show +/- 1 standard error of the mean.

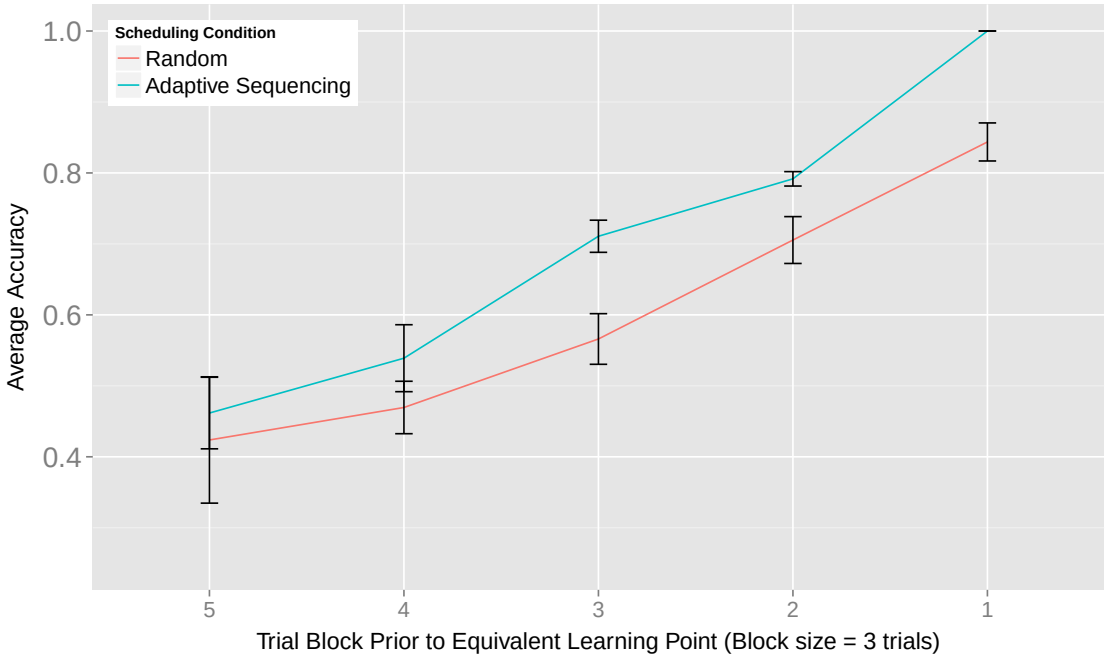

Figure 3.5: Exp 1. Accuracy at equivalent learning points for the last 3 presentations of each item in both scheduling conditions. Trial blocks were 3 presentations of each item. Error bars show +/- 1 standard error of the mean.

was conducted (trial block 5 was not included owing to that some participants did not have a 5th trial block [2]). The ANOVA found main effects of scheduling condition ($F(1,46)=11.9$, $p<.01$) and trial block ($F(3,138)=128.5$, $p<.001$), but no condition by trial block interaction ($F(3,138)=1.88$, $p=.14$). Independent t-tests were conducted at each trial block. Accuracies were higher for the adaptive condition than the random condition at blocks 1, 2, and 3 ($ts(46)=5.83$, 2.5, and 3.43 respectively, all $ps<.05$) but not at blocks 4 and 5 (block 4, $t(46)=1.16$, $p=.25$; block 5, $t(35)=0.354$, $p=.73$).

Accuracy change between pre-test and post-tests Finally, we also compared raw accuracies between conditions using a change score between pre and post-tests (Post-test accuracy minus pre-test accuracy). Accuracy change scores are shown in Figure 3.6 for immediate and delayed post-test change scores. A 2X2 mixed factor ANOVA on scheduling condition and post-test phase showed a significant main effect of condition ($F(1,46)=17.8$, $p<.001$), a main effect of post-test phase ($F(1,46)=105.5$, $p<.001$), and a significant test phase by condition interaction ($F(1,46)=10.9$, $p=.001$). At immediate post-test change scores were higher for the random condition ($M=0.93$, $SD=0.09$) than for the adaptive condition ($M=0.85$, $SD=0.12$), a significant difference ($t(46)=2.55$, $p=.01$, $d=0.746$). At delayed post-test, change scores were also higher for the random condition ($M=0.76$, $SD=0.16$), than for the adaptive condition ($M=0.52$, $SD=0.22$), a significant difference ($t(46)=4.29$, $p<.001$, $d=2.68$). Comparing conditions across test phases, the difference between immediate and delayed change scores for the random condition was significant ($t(23)=5.38$, $p<.001$) as was the difference for the adaptive condition ($t(23)=8.9$, $p<.001$).

3.3.3 Discussion

We found significantly greater learning efficiency in an adaptive condition where learners had trials scheduled using an adaptive scheduling algorithm that utilized learning to criterion and dropout, than in a random presentation condition that had no adaptive features, either for

[2]Some participants retired all items in less than 12 presentations; 3 trials x 4 blocks=12 presentations

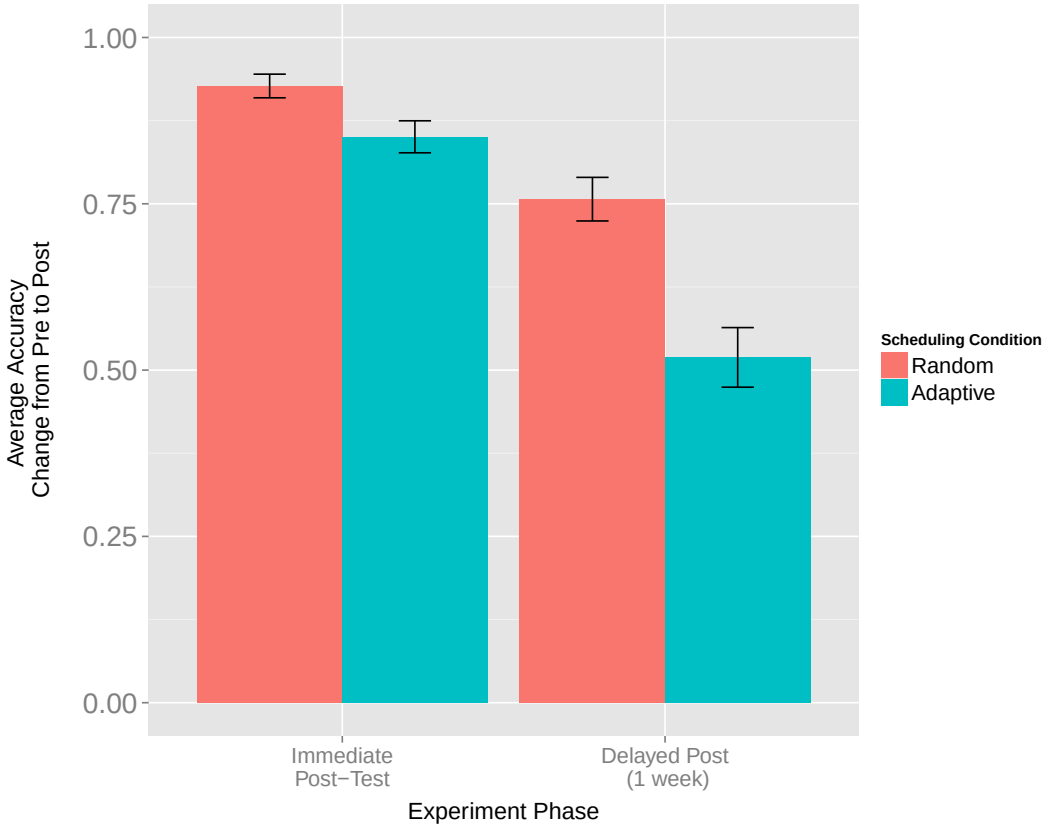

Figure 3.6: Exp 1. Accuracy change score (Post-tests minus pre-test accuracy) for immediate and delayed post-tests by scheduling condition. Error bars show +/- 1 standard error of the mean.

scheduling or dropout. Efficiencies - average learning gains weighted by the number of trials in a learning session - were higher in the adaptive condition for both immediate post-tests and delayed post-tests. Efficiencies were 102 percent higher for adaptive then random at immediate post-test and 50 percent higher at delayed post-test.

Thus, in sessions that had longer duration and used rules to determine when proficient learning had been met for each item, the adaptive algorithm succeeded in creating efficient, durable learning. This was despite a degree of overlearning that occurred in the random schedules - despite nearly twice as many presentations of each item, and despite overall higher accuracies in the random condition.

In addition, when compared at the same moment during learning, average accuracies were higher in the adaptive condition at a number of equivalent points of learning. This is further evidence that random schedules were simply less efficient than adaptive ones at creating successful learning, and some evidence indicating that extra trials in the random case were not unnecessary but actually improved learning. In addition to accuracy results, we also found lower reaction times in the adaptive condition indicating greater learning strength at retirement.

We were also interested in the effect of dropout on non-adaptive conditions of practice, and in a second experiment we compared adaptive sequencing to a random schedule of practice that also included drop-out. In Experiment 2, both the adaptive and random presentation conditions had exactly the same retirement criteria, the same as the adaptive condition in Experiment 1. That is, in Experiment 2, both adaptive and random conditions had learning criteria and dropout of individual items as each item met a learning criteria. This experiment would also provide evidence that it is not learning criteria alone that give adaptive schedules their power.

It was expected that adaptive conditions with learning criteria would still be better than random presentation conditions, even random presentation conditions with the same learning criteria.

3.4 Experiment 2: Random vs. Adaptive Presentation (Random With Dropout)

3.4.1 Method

Participants Participants were 48 UCLA undergraduates, some who participated for course credit and some who were recruited and paid 16 dollars for their time.

Design The design was identical to Experiment 1, except that the random condition was altered to have dropout; after each item reached a learning criterion it was removed from the set of learning items. The learning criteria was five of the last five presentations of each item correct with response times less than seven seconds, the same as in Experiment 1.

Materials The materials were identical to Experiment 1.

Procedure The procedure was identical to Experiment 1.

3.4.2 Results

Primary Dependent Measure: Learning Efficiency Results for efficiency at immediate and delayed post-tests are shown in Figure 3.7. A 2x2 mixed factor ANOVA on scheduling condition (adaptive vs. random) and post-test phase (Immediate vs. Delayed) found significant main effects of condition ($F(1,46)=10.6$, $p=.002$), a main effect of post-test phase ($F(1,46)=163.28$, $p<.001$), but no significant scheduling condition by test phase interaction ($F(1,46)=1.83$, $p=.18$). At immediate post-test, efficiencies were higher in the adaptive condition ($M=0.12$, $SD=0.02$) than the random condition ($M=0.09$, $SD=0.02$) a significant difference ($t(46)=4.07$, $p<.001$, Cohen's $d=1.17$). At delayed post-test, efficiencies were also higher in the adaptive condition ($M=0.084$, $SD=0.03$) than the random condition ($M=0.064$, $SD=0.064$), a significant difference ($t(46)=2.31$, $p=.025$, Cohen's $d=0.67$). Comparing means between the two test phases, the difference between efficiencies at each test phase for both the adaptive and random condition were significant (adaptive immediate vs. delayed, $t(23)=9.05$, $p<.001$; random immediate vs. delayed, $t(23)=9.14$, $p<.001$).

Trials to criterion Despite identical drop-out features, learning took differing amounts of time in the two scheduling conditions. The number of trials to criterion was analyzed. The number of trials to criterion in each condition are shown in Figure 3.8. The random condition took on average 231 trials to reach the end of the session ($SD=48.3$). The adaptive

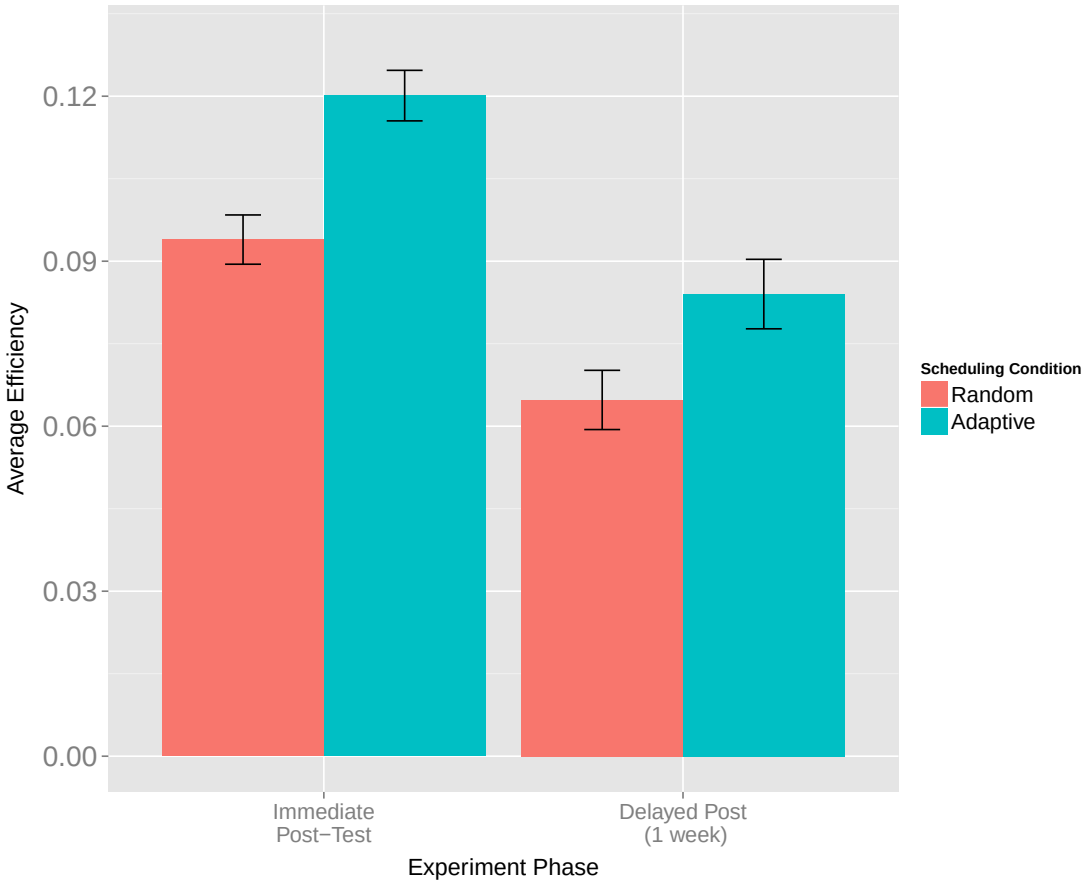

Figure 3.7: Exp 2. Efficiency at immediate and delayed post-tests by scheduling condition. Efficiency was the change in accuracy from pre-test to post-test divided by the number of trials in training. Error bars show +/- 1 standard error of the mean.

condition took on average 183 trials (SD=35.1). This difference was significant ($t(46)=3.88$, $p<.001$, $d=1.13$).

Equivalent learning trials analysis Because schedule structure led to differing numbers of trials during training, we analyzed the two scheduling conditions at equivalent points near the end of training - at the moment that participants in the adaptive condition retired - in order to assess whether learning was still progressing in the random condition. Comparing at equivalent moments in learning allows for an alternative comparison of training performance than average accuracy at a post-test, where, owing to learning criteria, performance

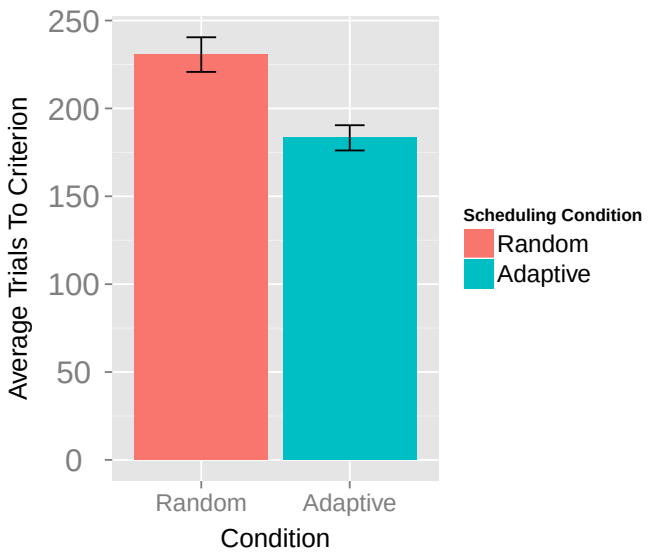

Figure 3.8: Exp 2. Number of trials in learning session for each scheduling condition. Error bars show +/- 1 standard error of the mean.

is expected to be the same across conditions.

Accuracies were compared at the moment participants in the adaptive condition finished the learning phase, trial 183. The accuracy of the prior 2 presentations for each item was measured in both conditions. Accuracies are shown in Figure 3.3. Average accuracy was higher in the adaptive condition than the random condition (adaptive M=1.0, SD=0; random M=0.91, SD=0.13), a significant difference (t(46)=3.46, p=.001, d=1.41).

Average reaction times were also measured at the equivalent point in learning. Reaction times are shown in Figure 3.10. Reaction times were lower for the adaptive condition (M=2.66 sec, SD=0.33) than the random condition (M=3.36 sec, SD=0.86), a significant difference (t(46)=3.22, p=.002, d=1.02).

We also analyzed accuracies at a number of trial blocks before the end of the learning phase (trial 183) in the adaptive condition. At five equivalent learning points, average accuracies for the last 3 trials per item are displayed in Figure 3.11 (blocks are displayed in reverse in the figure). A 2X3 mixed factor ANOVA on schedule condition and trial block

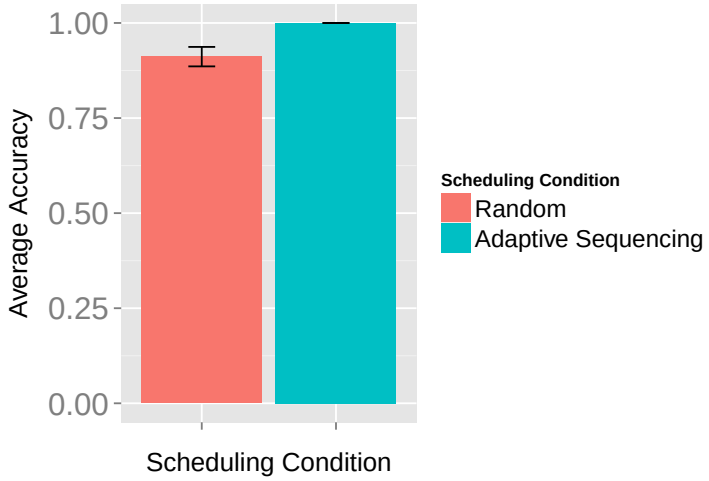

Figure 3.9: Exp 2. Accuracy at equivalent points in learning (trial 183) in both scheduling conditions. Error bars show +/- 1 standard error of the mean.

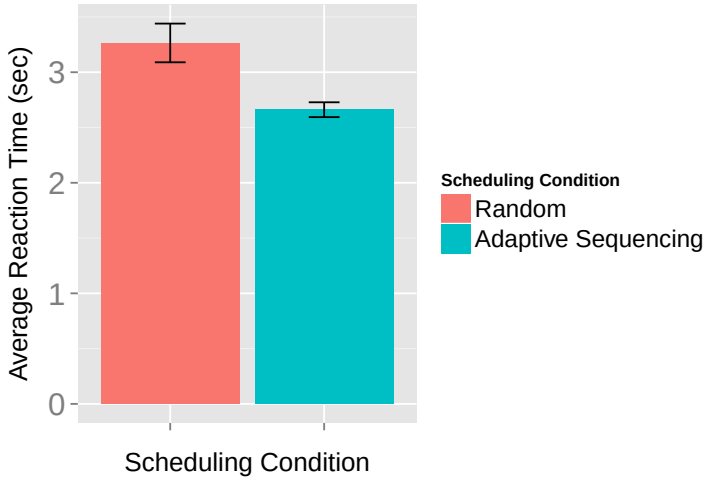

Figure 3.10: Exp 2. Reaction time at equivalent points in learning (trial 183) in both scheduling conditions. RTs are from correct responses only. Error bars show +/- 1 standard error of the mean.

was conducted (trial blocks 4 and 5 were not included owing to that some participants did not have a complete 4th or 5th trial block [3].). The ANOVA found main effects of scheduling condition ($F(1,46)=45.1$, $p<.001$) and trial block ($F(2,92)=285.7$, $p<.001$), and a

[3]Some participants retired all items in less than 9 presentations; 3 trials x 3 blocks=9 presentations

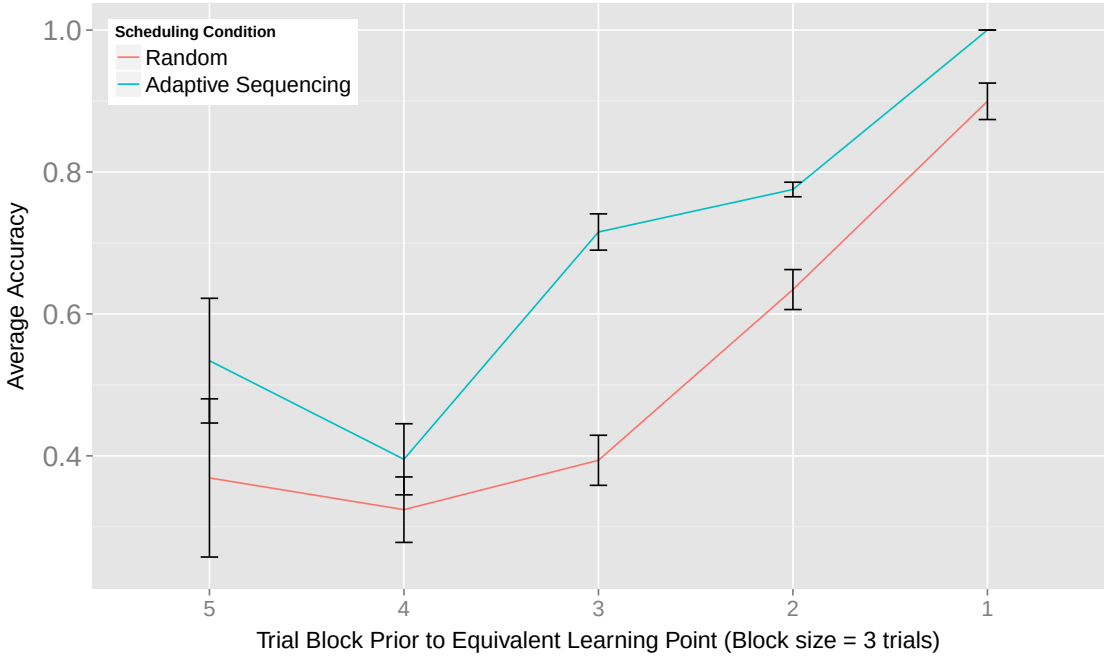

Figure 3.11: Exp 2. Accuracy at equivalent learning points for the last 3 presentations of each item in both scheduling conditions. Trial blocks were 3 presentations of each item. Error bars show +/- 1 standard error of the mean.

significant condition by trial block interaction ($F(2,92)=24.9$, $p<.001$). Independent t-tests were conducted at each trial block. Accuracies were higher for the adaptive condition than the random condition at blocks 1, 2, and 3 ($ts(46)=3.91$, 4.71, and 7.38 respectively, all $ps<.001$) but not at blocks 4 and 5 (block 4, $t(45)=1.04$, $p=.30$; block 5, $t(26)=1.14$, $p=.26$).

Accuracy change between pre-test and post-tests Finally, we also compared raw accuracies between conditions using a change score between pre and post-tests (Post-test accuracy minus pre-test accuracy). Accuracy change scores are shown in Figure 3.12 for immediate and delayed post-test change scores. A 2X2 mixed factor ANOVA on scheduling condition and post-test phase showed no significant main effect of condition ($F(1,46)=0.11$, $p=.75$), a significant main effect of post-test phase ($F(1,46)=160.72$, $p<.001$), and no significant test phase by condition interaction ($F(1,46)=0.01$, $p=.94$). At immediate post-test change scores were higher for the adaptive condition ($M=0.89$, $SD=0.07$) than for the random condition

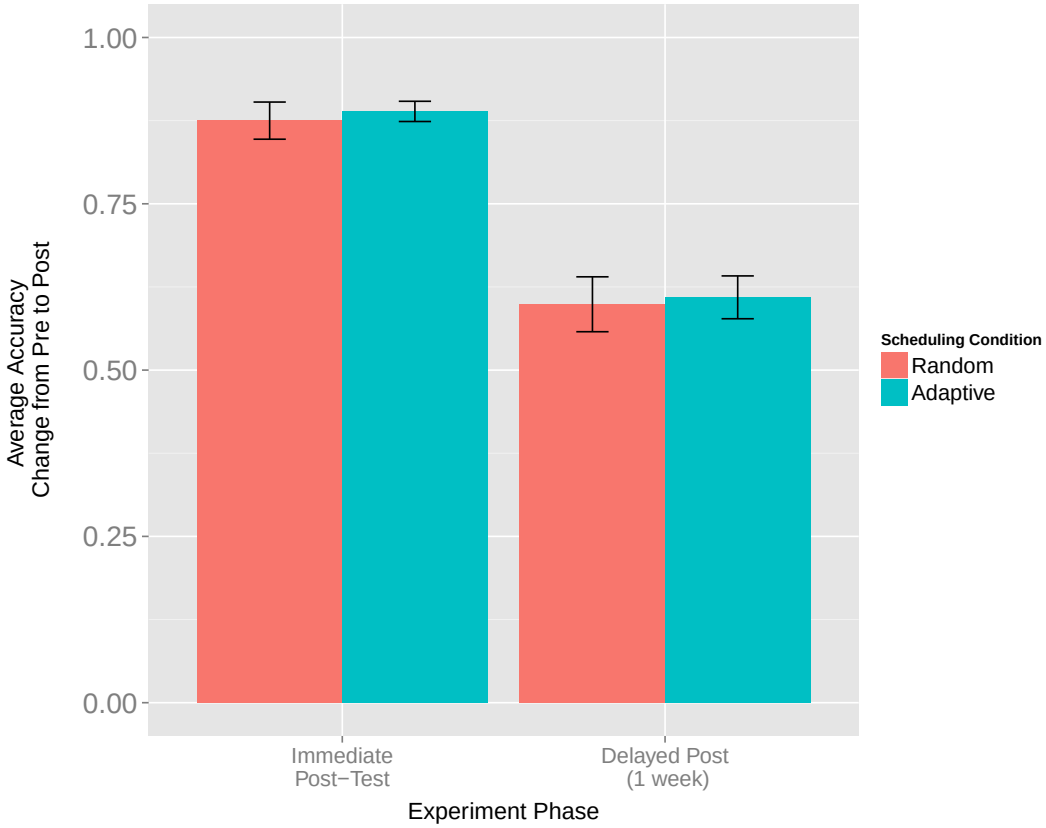

Figure 3.12: Exp 2. Accuracy change score (Post-tests minus pre-test accuracy) for immediate and delayed post-tests by scheduling condition. Error bars show +/- 1 standard error of the mean.

(M=0.87, SD=0.14), not a significant difference (t(46)=0.44, p=.66, d=0.132). At delayed post-test, change scores were also higher for the adaptive condition (M=0.61, SD=0.16), than for the random condition (M=0.60, SD=0.20), also not a significant difference (t(46)=0.20, p=.84, d=1.17). Comparing conditions across test phases, the difference between immediate and delayed change scores for the random condition was significant (t(23)=9.18, p<.001) as was the difference for the adaptive condition (t(23)=8.77, p<.001).

3.4.3 Comparisons between Experiment 1 and 2

We also conducted one short analysis of the differences between learning efficiency in Experiments 1 and 2. Specifically we compared the effects of random schedule presentation without dropout (Exp. 1) and with dropout (Exp 2.) to approximate the effects of dropout on efficient learning. The difference between random conditions was significant ($t(46)=3.2$, $p=.002$, Cohen's $d=0.924$).

3.4.4 Discussion

Just as in Experiment 1, average efficiencies were higher for the participants who had learning trials scheduled using an adaptive algorithm than for participants who had learning trials scheduled randomly. This was true even though both schedules possessed exactly the same retirement criteria (including both accuracy and reaction time criteria). The efficiencies in the adaptive condition were 33 percent higher than the random condition at immediate post-test, and 31 percent higher for the adaptive condition than the random condition at delayed post-test.

In addition, an analysis of learning session accuracies at equivalent points in training found that learners were on average more accurate in the adaptive condition than in the random condition, both at an equivalent point and for some trial blocks previous to that point.

3.5 Conclusion

3.5.1 Discussion of Adaptive Sequencing and Learning to Criterion

Efficiencies improved for schedules that utilized adaptive sequencing over a random schedule of practice. In both experiments, adaptive sequencing included dropout, or removal of items when the learning for an item reached a criterion level of performance - namely, five of the last five presentations of the item had to be answered correctly and in under 7 seconds. In

Experiment 1, random schedules did not include dropout. This comparison demonstrated the power of adaptive sequencing to create schedules of practice that are more beneficial than what might be typically accomplished in a real learning situation. In Experiment 2, random schedules included dropout. This comparison showed that adaptive schedules were more efficient than random schedules, even when equated for dropout. Clearly, the structure of adaptive trial schedules drove efficient learning, and dropout alone was not the driver of learning effects in adaptive sequencing.

In addition, dropout improved the efficiency of schedules. When comparing the efficiency of random schedules without dropout (Experiment 1) to random schedules with dropout (Experiment 2), learning efficiencies were greater for schedules that included dropout. Analytically, efficiencies should increase with the removal of items - schedules will use fewer trials if items drop out - but dropout alone does not determine final learning efficiency. Dropout must be combined with efficient spacing schedules that encourage robust learning.

We have demonstrated that, when comparing longer schedules of practice that continue to learning criteria, adaptive schedules outperform random schedules of practice - where random schedules are effective but under-studied forms of scheduled practice. Adaptive schedules were more efficient - that is, the rate of learning and the durability of that learning were greater - than random schedules, both at delayed tests as well as at equivalent points during the learning session.

Other questions There were a number of theoretical reasons that we expected adaptive scheduling to perform better than fixed schedules of practice. One hypothesis was that adaptive schedules benefit learning because they better account for random fluctuation in the learning process; meaning that the match between adaptive schedules and actual learning strength will be high, whereas fixed schedules, especially fixed equal and expanding schedules, will deviate from actual, ongoing changes in learning strength and produce poorer learning gains. Although random schedules too should benefit from random fluctuations in learning since random schedules by definition fluctuate, those fluctuations in random conditions rarely

would meet the requirements of ongoing learning strength of individual items and learners. Thus, their performance may very infrequently resemble that of adaptive schedules, but it appears the vast majority of the time, adaptive schedules better match ongoing learning strength.

These studies by no means represent the last word on the issues discussed here. The relationship between learning criteria, spacing and learning performance will prove to be a fluid one; no one factor can be examined in the absence of others. Increases in the strictness of learning criteria will interact with spacing; and the benefits to efficiency brought by dropout will be balanced by the efficiencies brought by the structure of learning schedules.

CHAPTER 4

Perceptual Learning and Adaptive Category Sequencing

Enhancing Complex Knowledge Representations

4.1 Introduction

4.1.1 Perceptual Learning and the Development of Expertise

Attaining expertise in many domains depends on not just the storage of individual facts, but changes in the way information is extracted – a process known as perceptual learning (Gibson, 1969; Kellman & Garrigan, 2009). In the last two decades, work in the cognitive and neural sciences has witnessed a resurgence of interest in perceptual learning (PL). The focus of early PL research (Gibson, 1969) and the application of PL in virtually all real-world tasks involves discovery of invariance amidst variation. As emphasized by Gibson, perceptual learning is a process of picking up relevant structure and ignoring irrelevant structure; of discriminating between highly similar categories. Gibson often used "differentiation learning" as a synonym for PL. In this process, effects such as pattern recognition, discovery, transfer, and fluency come to the fore. These are neural and information processes of a potentially different character than the mnemonic processes of encoding and recall.

Whereas item learning involves storing and retaining specific information, PL has been argued to contain two kinds of changes: discovery and fluency effects (Kellman, 2002). Discovery involves altering encoding processes to progressively locate the most relevant in-

formation for some task. Specific PL discovery effects, observed in both simple and complex PL tasks, include increasing selectivity and precision of information extraction as learning progresses; relevant features are encoded and irrelevant ones ignored (Gibson, 1969; Petrov, Dosher, & Lu, 2005). Other discovery effects involve the learner coming to notice higher-order relations that were initially not encoded at all, and/or coming to encode information in larger "chunks" (Chase & Simon, 1975; Gibson, 1969; Goldstone, 2000; Kellman & Garrigan, 2009). Fluency effects involve improved speed, greater parallel processing, and lower attentional load in picking up task-relevant information as learning progresses (Kellman & Garrigan, 2009; Schneider & Shiffrin, 1974).

All of these processes go well beyond storing and maintaining a specific memory trace. In learning a number of related perceptual classifications, commonalities or invariances that determine category membership must be discovered in PL, and conversely, in learning to differentiate different categories, distinguishing features must be discovered.

The realization of the importance of PL in diverse learning tasks and the emergence of PL interventions raise the question of whether PL shares principles that have been found to improve or optimize other kinds of learning. When we learn new perceptual classifications, what principles govern successful learning? Are there ways of organizing the order of presentation of material such that learning is enhanced? Such questions form the basis for the following studies, which investigate effective training strategies for enhancing perceptual learning – especially when learning concerns sets of categories or natural kinds.

4.1.2 Can Scheduling Benefit Perceptual Category Learning?

Despite differences in underlying mechanism, there are reasons to suspect that spacing may be beneficial in perceptual category learning as well as in factual learning. One reason is that interleaving exemplars of different categories may facilitate discovery of distinguishing features (Gibson, 1969), just as paired comparisons might (Mettler & Kellman, 2009; Kang & Pashler, 2012). On the other hand, discovering perceptual attributes shared by exemplars

of a single category might better be served by encountering several exemplars close together in time; in other words, massed rather than spaced presentation. Perhaps a more compelling reason for an analogy between spacing benefits in fact-learning and PL is the notion that the best time to receive further practice is when learning strength has declined enough to make accurate performance relatively difficult. Although different specific mechanisms of learning may be at work in different domains, optimizing practice based on intervals that progressively increase as learning strength grows may be a commonality across types of learning.

There has been some earlier work on these questions. Kornell and Bjork (2008) compared interleaving and massing of learning items in perceptual category learning of artist's painting styles. In the interleaved condition, one painting from each artist was presented in a sequence before any second painting from an artist was presented (each block of presentations contained 1 painting from each artist). In the massed condition, the 6 examples of each artist's paintings were presented consecutively, followed by the entire set of another artist's paintings, and so on, until all paintings from all artists were presented. They measured participants' accuracy in classifying previously unseen paintings from each artist and found that interleaving led to greater success.

Kornell and Bjork's results differ from that of some studies in memory and human performance that show advantages for blocked vs. randomized trials of practice (see Schmidt & Bjork, 1992, for a review). Similarly, some work in perceptual learning and unsupervised category learning shows benefits for massing, severe detriments for interleaving of stimuli (Zeithamova & Maddox, 2009; Kuai, Zhang, Klein, Levi, & Yu, 2005), or no advantage for either type of schedule (Carvalho & Goldstone, 2011). We wondered if schedules that combine spacing with modest amounts of massing could result in even greater learning than spacing or massing alone. This hypothesis was tested using a separate condition that combined blocking in the initial stages of learning with adaptive spacing in later stages (see below).

4.1.3 Current work

To study this question, we used a learning task involving taxonomic classifications of images of butterfly (Lepidoptera) species. Natural stimuli such as these afford the type of feature discovery present in real-world perceptual learning, where, in contrast to most artificial stimuli, relevant stimulus features are richly perceptual, hierarchically organized, and distributed stochastically and non-independently across categories. We employed a web-based perceptual learning module (PLM) that included the ARTS system described in prior Chapters. The PLM presented butterfly images in pairs, one from a target category (target butterfly genus) and one from an alternate category. Participants were asked to choose the image that correctly matched the presented target category name. Feedback was given on each trial, and participants continued discriminating butterflies until they had learned the correct label-to-category mappings. Previous work suggests that paired comparisons across many trials are effective in eliciting PL (Mettler & Kellman, 2009; Walheim, Dunlosky, & Jacoby, 2011).

We continued the PLM until each learner met mastery (learning) criteria based on accuracy and speed of classification. We used learning criteria because of their relevance to applications in real-world learning contexts, and from the standpoint of experimental control, they allowed us to assess learning after each learner had reached a similar endpoint. Learning criteria are also used to determine dropout of learned categories, allowing each learner to spend further learning effort where it is needed most. The benefits of using learning criteria, however, come with a difficulty. Different learners require different numbers of trials to reach criterion. This leaves the experimenter with two dependent measures of learning – post-test performance and trials to criterion. To allow comparisons between conditions that included both of these measures, we combined them into a measure of learning rate or learning efficiency, as discussed in prior Chapters, defined as accuracy gains divided by learning trials invested.

Based on potential benefits of both spacing and massing in PL, and in view of earlier

work indicating that complete massing is sub-optimal, we also included a condition with some initial massing of category exemplars ("mini-blocks"). To ensure that learning involved discovery of perceptual structure, rather than memorization of instances, we assessed learning using an equal number of unfamiliar instances, never shown during the pre-test or learning phases, and familiar instances, which could appear one or more times during learning. Finally, we tested learning both in an immediate post-test and after a delay of one week. In studying spacing effects in perceptual learning, it is important to consider the possibility of transient performance effects that appear in immediate tests but might not survive in a delayed test, as has proven important in studies of other kinds of learning. Conditions that optimize performance on immediate tests may not be the ones that are best for durable learning (e.g., Schmidt & Bjork, 1992).

In Experiment 1a we tested three conditions: 1) a control condition that orders items in an unmodified random sequence, 2) an adaptive sequencing condition (ARTS) that changes the delay between presentations of a category as a function of learning strength, and 3) a condition that initially blocks 3 exemplars from a category sequentially (called "mini-blocks"), presented for two rounds before proceeding to standard adaptive sequencing. Because the adaptive sequencing conditions also included retirement, we anticipated that learning would be quicker there, and that performance at an immediate and 1 week delayed test would be most efficient for participants in those conditions (greatest learning per trial invested in training).

In a second experiment (1b), we manipulated the degree of variability between categories to observe how adaptive sequencing interacts with category structure. Decreasing the variability, and thus making items within a category more similar to one another, is a way of approximating the effect that dynamic sequencing would have on a variety of types of categories. It is also more similar to a situation in which learning concerns individual items as opposed to categories of varying exemplars. Our prediction was that adaptive sequencing would operate as well when categories were of lower variability as when high, and we tested the efficacy of the adaptive scheduling algorithm in both cases without any modification to

the parameters of our model.

4.2 Experiment 1a: Perceptual Category Sequencing vs. Random Presentation

4.2.1 Methods

In perceptual category learning using an adaptive sequencing algorithm, all trials consist of a two-alternative forced choice procedure between two images. Participants are asked to choose which image matches a presented category label. Feedback is presented on each trial.

Participants 54 undergraduate psychology students from the University of California, Los Angeles participated in an hour-long experiment for course credit. Participants returned one week after the first session for a delayed post-test. Four of 58 were disqualified: one because of a failure to complete a delayed post-test and three others, one from each condition, who failed to reach a learning criterion in the training session, as described below.

Materials The materials for this study consisted of 108 images of Lepidoptera (butterfly) specimens arranged into 12 categories by genus (See Figure 4.1 for examples). Each

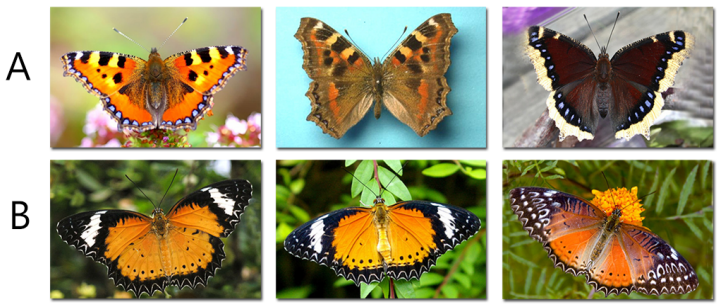

Figure 4.1: Exp. 1a & 1b. Examples of butterfly images used in the experiments. Three examples from each of two butterfly genera (trained categories) are shown. A) Examples of genus Aglais. B) Examples of genus Cethosia.

category contained nine exemplars where one exemplar from each category was withheld during learning in order to be used as a test of transfer of learning to unseen items in the two post-test phases (see Figure 4.2). A multidimensional scaling analysis was conducted

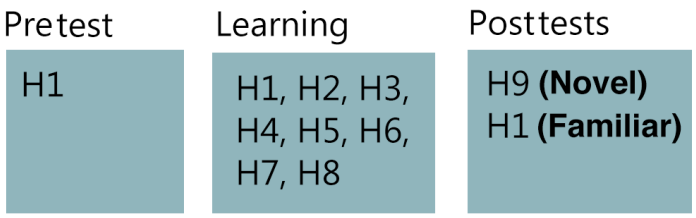

Figure 4.2: Exp 1a & 1b. Example distribution of one stimulus category across experiment phases. In pre-test: Exemplar H1 is tested. In the learning phase, all category exemplars except H9 are presented. In each post-test, one previously seen exemplar, H1, and one novel exemplar not presented during the learning phase, H9, are tested.

to ensure categories occupied positions of roughly equivalent similarity distance from each other (implying equivalent learning difficulty), though there was some variability in difficulty across categories. The images used for both the immediate and delayed post-tests were fixed for each subject. Images were presented in jpeg form in 16-bit color, where each image was 450x300 pixels. All pre-test, training, and post-test sessions occurred within a web-based perceptual learning module (PLM). The PLM presented a text label of the category name in an upper middle position (as in the 'sample' position, of a 'match-to-sample' presentation). In pre-test and post-test trials, four images were presented in the center of the screen in two rows and two columns (see Figure 4.3a). Only one image contained an exemplar from the target category – the distractors each contained an exemplar from one of three alternate categories. During training trials, two images were shown side by side in the middle of the screen just below the category label (a 2AFC presentation, see Figure 4.3b).

Design The experiment utilized a pre-test/post-test design. A pre-test measured baseline levels of perceptual category knowledge. Participants completed 12 trials where each category

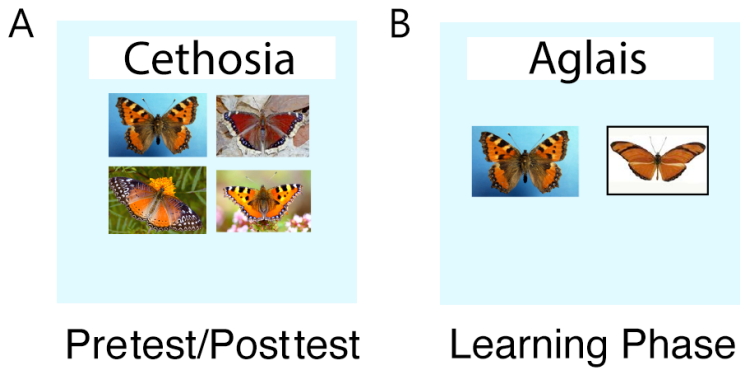

Figure 4.3: Exp 1a & 1b. Trial presentation formats in the assessment and learning phases of the experiments. A) Pre-test and post-test: Each trial was a 4-alternative forced choice, where one of the 4 exemplars belonged to the target genus. B) Learning phase: Each trial was a 2-alternative forced choice, where one of the two exemplars belonged to the target genus.

was presented as a target once, in random order. Each trial consisted of a match-to-label test; a four alternative forced choice between four images: exemplars from three incorrect categories and one exemplar from the correct target category. Pre-test exemplars were randomly chosen at the start of the experiment and the same exemplars were displayed to all participants.

The training session consisted of a series of match-to-label trials, where each trial tested one target category. Trials consisted of a two alternative forced choice (2AFC) decision between two images: a randomly selected exemplar from the target category and a randomly selected exemplar from an alternate category. There were 3 between-subjects scheduling conditions that determined the order of presented categories: 1) purely random stimulus presentation 2) adaptive category sequencing with retirement, and 3) mini-blocks with adaptive category sequencing and retirement. The participant completed as many trials as necessary to reach a learning criterion.

An immediate post-test measured the degree of perceptual learning after a learning session. The immediate post-test was similar to the pre-test but contained an additional trial

for each category, for a total of 24 test trials. For each category, one trial was of a familiar exemplar (an image presented during training) and one trial was of a novel exemplar (not presented during training). The novel exemplar was used to measure transfer or generalization of category knowledge to unseen stimuli. The same post-test items were shown to all participants. A delayed post-test, given one week later, was identical to the immediate post-test and measured the amount of retention after a delay.

Adaptive sequencing algorithm The adaptive sequencing algorithm was identical to that presented in Chapters 1-3 except that some sequencing parameters were different ('RT weight', r=1.7; 'enforced delay', D=2) and the algorithm assigned priority to categories of stimuli rather than to individual items.

In all conditions, when a category was chosen for presentation, items were randomly chosen exemplars from the target category, where the odds of any exemplar being selected were 1/8.

Other scheduling conditions Adaptive sequencing with 'mini-blocks' (Adaptive/Mini-blocks condition) was identical to the Adaptive condition, but at the start of training participants received 'mini-blocks' of 3 exemplars from the same category consecutively presented across sequential trials. Participants received two 'mini-blocks' per category before adaptively sequencing individual presentations of categories without blocking. We hypothesized that in this condition, a moderate degree of grouping of exemplars would aid in comparison processes known to enhance perceptual learning and category learning.

In the Random presentation condition, training sessions consisted of random selection of categories on each trial, with no constraints on the total number of times a category could be presented or the total number of presented stimuli from each category. This condition implemented a method for ending training after the accuracy for every category had reached the same retirement criteria as in the dynamic sequencing condition (5 out of 6 correct). This helped to ensure that the number of total presentations of individual categories would accu-

rately reflect typical randomized learning schedules and remain distinct from the category retirement feature that was present in adaptively sequenced schedules.

Procedure In the pre-test, participants were presented with a category label at the top of the screen and four images in the center of the screen. Participants were instructed to indicate the image that belonged to the presented category label and to make their best guess if they did not know the answer. No feedback was given during this phase and the test took no more than 3-5 minutes. The learning phase consisted of one session, no longer than 45 minutes, where participants were instructed to choose the image that best matched a presented genus label. Participants were shown one genus name at the top of the screen and images from two different butterfly genera side by side. Participants were asked to choose either the left or right image and respond using the keyboard. Responses were considered correct if the chosen image belonged to the correct genus. Participants were given 30 seconds to respond and were always provided with feedback. If a participant failed to respond within 30 seconds, the trial timed out and feedback was given, where a timeout was recorded as an incorrect response. Feedback consisted of highlighting the correct image and displaying "correct" or "incorrect" depending upon the accuracy of the participant's response. In addition, the name of the target genus moved to a position underneath the correct image. Feedback displayed for a minimum of 3.5 seconds, although participants had up to 15 seconds to view the feedback before the screen was cleared. Participants could use the spacebar to progress to the next trial any time after the initial 3.5 seconds. Summary feedback was provided every 10 trials. Summary feedback consisted of a graph of average accuracies and response times for each previous 10 trial block. Immediately after training, participants completed a post-test. After the post-test participants were asked to not study or review the information in the study. One week after the post-test, a delayed post-test was administered.

Dependent measures and analyses We used an efficiency measure as in Experiments 1 and 2 in Chapter 3. We were also able extract a measure of learning after equal numbers of learning trials. To do this, we took the average number of learning trials required in the basic

adaptive condition to reach the learning criteria, and we looked at accuracy and response times for conditions on the last two presentations of each stimulus category at that point in learning. This measure allows some indication of learning across conditions at a point where each condition had the same number of learning trials.

Predictions In the Adaptive condition, it was expected that adaptive sequencing – where quick, correct answers to categories would delay their reappearance – would lead to more rapid learning and enhance discrimination even for difficult categories. It was thought that retirement of well learned categories would make learning more efficient to an even greater degree. We expected that the partial blocking in the Adaptive/Mini-blocks condition would perform better than the Random condition and that the Adaptive/Mini-blocks condition might even outperform the basic Adaptive condition. A similar effect of these conditions on transfer to novel stimuli was expected.

4.2.2 Results

Learning performance was measured using a pre-test, a post-test administered immediately after training, and a delayed post-test administered one week after training. Pre-test scores averaged 2.43 items out of 12, indicating that performance was no better than chance and that participants did not possess prior knowledge of butterfly genera. A between subjects ANOVA on proportion correct in the pre-test confirmed no significant differences across the three conditions ($F(2,51)=1.86$, $p=.17$, $\eta^2=0.07$). Individual comparisons also showed no reliable differences (all ps $>$ .20 for Random vs. Adaptive, Random vs. Adaptive/Mini-blocks, Adaptive vs. Adaptive/Mini-blocks, respectively). In addition, an examination of participant reaction times (RTs) on the pre-test showed no reliable difference between conditions, neither in a between subjects ANOVA nor in individual comparisons (all ps $>$ .19).

Learning efficiency was measured in the immediate and delayed post-tests by dividing the number of correct post-test items by learning trials invested. Efficiency scores are shown in Figure 4.4 for the three scheduling conditions, and for both previously seen and novel

instances in both immediate and delayed post-tests. Efficiencies for the Adaptive condition

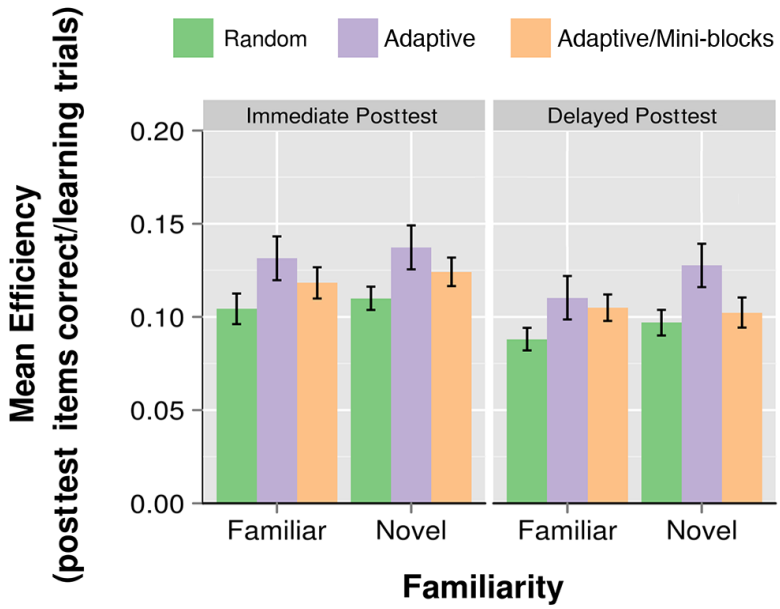

Figure 4.4: Exp 1a. Mean efficiency scores by learning condition and post-test phase. Efficiency scores were the number of post-test items answered correctly divided by the number of trials invested in learning. Familiar stimuli were post-test items that had been shown during training, whereas novel stimuli were items that had not been presented previously. Error bars indicate +/- one standard error of the mean.

were numerically higher than efficiencies in the Random and Adaptive/Mini-blocks condition for both immediate and delayed tests (Immediate post-test: M=0.13, vs. 0.11 and 0.12; Delayed post-test: M=0.12 vs. 0.09 and 0.10 respectively). We performed a 3 (condition – Adaptive, Random and Adaptive/Mini-blocks) by 2 (post-test phase – immediate vs. delayed) by 2 (previously seen vs. novel) mixed factor ANOVA with condition as a between-subjects factor and test phase and stimulus familiarity as within-subjects factors. The ANOVA revealed a marginally reliable main effect of scheduling condition ($F(2,51)=2.45$, $p=.096$, $\eta^2=0.09$). There was a reliable main effect of test phase ($F(1,51)=51.52$, $p < .001$, $\eta^2=0.50$), and no interaction of scheduling condition with test-phase ($F(2,51)=0.16$, $p=.86$,

η^2=0.006).

There was a strong main effect of stimulus familiarity (F(1,51)=17, p < .001,η^2=0.25), due to the somewhat surprising result that performance in the post-tests was superior for novel instances. There was also a marginally reliable interaction between condition and familiarity (F(2,51)=3, p=.059, η^2=0.105), apparently due to the greater superiority of transfer to novel instances in the Adaptive condition. There was no interaction between test phase and familiarity (F(1,51)=0.30, p=.59, η^2=0.006), nor was the three way interaction between condition, phase and familiarity reliable (F(2,51)=2.45, p=.097, η^2=0.08). Examining the interaction of condition and familiarity, there was a reliable efficiency advantage for the Adaptive vs. the Random condition for novel items at both immediate post-test (Adaptive: M=0.14; Random: M=0.11; t(34)=2.04, p < .05) and delayed test (Adaptive: M=0.12; Random: M=0.09; t(34)=2.27, p=.03) but the numerical advantage was marginal or unreliable for familiar items at immediate (Adaptive: M=0.13; Random: M=0.10; t(34)=1.69, p=.10) or delayed test (Adaptive: M=0.11; Random: M=0.09; t(34)=1.89, p=.067). Otherwise, there were no reliable differences between conditions with either novel or familiar stimuli at any test (all ps > .05). Because our hypothesis specifically concerned differences in conditions we conducted planned paired comparisons between each condition. Averaging across post-tests, t-tests showed that the difference between Random and Adaptive conditions was significant (t(34)=2.08, p < .05, Cohen's d=0.72). On the immediate post-test, these two conditions differed marginally (t(34)=2.02, p=.051, d=0.70) and on the delayed post-test, there was a reliable advantage for the Adaptive condition (t(34)=2.04, p < .05, d=0.71). Other comparisons between scheduling conditions were not significantly different (averaging over post-tests, Random vs. Adaptive/Mini-blocks, p=.20, d=0.36; Adaptive vs. Adaptive/Mini-blocks: p=.33, d=0.43). Paired t-tests showed a reliable decrease between immediate vs. delayed post-tests for all three conditions (all ps < .005).

The efficiency advantage for the Adaptive condition compared to the Random condition amounted to 25% in the immediate post-test and 29% in the delayed post-test.

We also analyzed separately the two dependent measures that were components of the

efficiency measure, number of learning trials and post-test accuracy for each participant. A between subjects ANOVA found significant differences between the number of training trials across the three schedules. Participants averaged 154.7, 167.4, and 204.3 trials in the Adaptive, Adaptive/Mini-blocks, and Random conditions, respectively - a reliable difference ($F(2,51)=5.50$, $p=.007$, $\eta^2=0.18$). Comparing means, the differences between the Adaptive and Random condition and between the Adaptive/Mini-blocks and Random condition were significant ($t(34)=3.02$, $p=.005$, Cohen's $d=1.01$ and $t(34)=2.42$, $p=.021$, Cohen's $d=0.81$ respectively). Trials did not differ reliably between the two adaptive conditions ($t(34)=.85$, $p=.04$, Cohen's $d=0.28$).

Raw accuracy data (not corrected for number of trials invested) are shown in Figure 4.5 for each condition in both immediate and delayed post-tests. A 3x2 ANOVA across scheduling conditions and both post-test phases found no effect of scheduling condition ($F(2,51)=1.94$, $p=.153$, $\eta^2=0.07$), an effect of test phase ($F(1,51)=52.8$, $p < .001$, $\eta^2=0.508$), and no interaction of phase and condition ($F(2,51)=0.053$, $p=.95$, $\eta^2=0.002$). Accuracies in the Random condition numerically exceeded those in the Adaptive and Adaptive/Mini-blocks conditions in both the immediate and delayed post-tests (Immediate: M: .86 vs. .79 & .80, respectively; Delayed: M: .75 vs. .68 & .69, respectively). Individual comparisons showed a marginally significant difference between the Random and Adaptive conditions on the immediate post-test ($t(34)=1.88$, $p=.069$, $d=0.63$); however, the difference was not reliable at delayed post-test ($t(34)=1.63$, $p=.11$, $d=0.55$). No reliable differences were found in immediate post-test accuracy between the Random and Adaptive/Mini-blocks condition or between the two Adaptive conditions ($ts(34)=1.57$ and 0.30, $ps=.13$ and .76, respectively). Similarly, at delayed post-test, there was no reliable difference between the Random and Adaptive/Mini-blocks conditions, or between the two Adaptive conditions ($ts(34)=1.34$ and 0.10, $ps=.19$ and .92, respectively). All three conditions showed accuracy decreases between post-test and delayed post-test (all $ps < .05$).

We carried out an additional accuracy analysis by comparing the three learning conditions at a point when all three had the same number of learning trials. We determined the mean

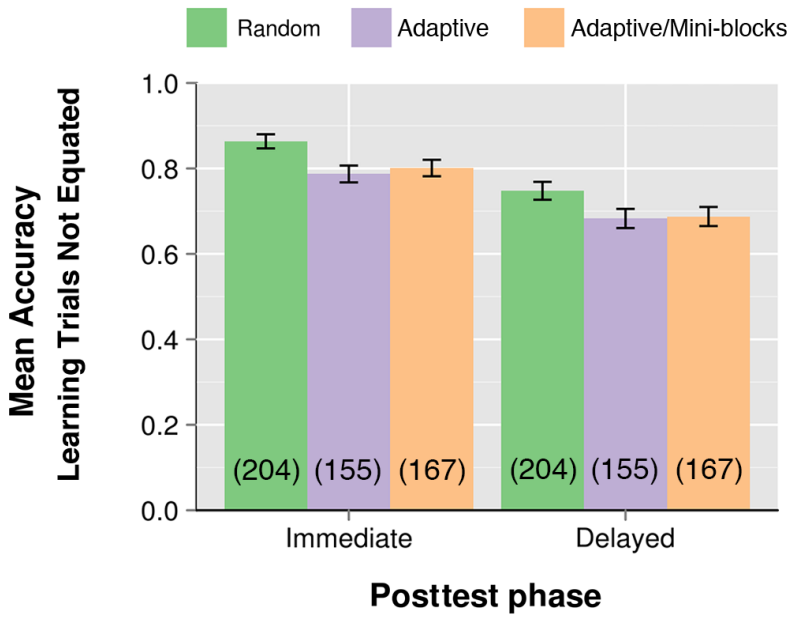

Figure 4.5: Exp 1a. Mean accuracy results by learning condition and post-test phase. Accuracy is given as the percentage of 24 post-test questions answered correctly. These data indicate raw accuracy not corrected for number of learning trials; the average number of learning trials in each condition is shown in parentheses. Error bars indicate +/- one standard error of the mean.

number of trials to reach criterion in the standard Adaptive condition and examined the performance of learners in the Random condition and the Adaptive/Mini-blocks condition after the same number of trials. The mean number of trials to reach learning criterion in the Adaptive condition was 155 trials (SD=48.2). In the Adaptive condition, we calculated the average accuracy across the last two presentations of each learning category at the time each learner reached learning criterion. In the Random and Adaptive/Mini-blocks conditions we calculated the average accuracy across the last two presentations of each learning category at the point when learners had received 155 learning trials. Mean proportions correct were .98, .96, and .92 for the Adaptive, Adaptive/Mini-blocks, and Random conditions respectively, after an average of 155 learning trials (see Figure 4.6). An ANOVA showed a reliable main

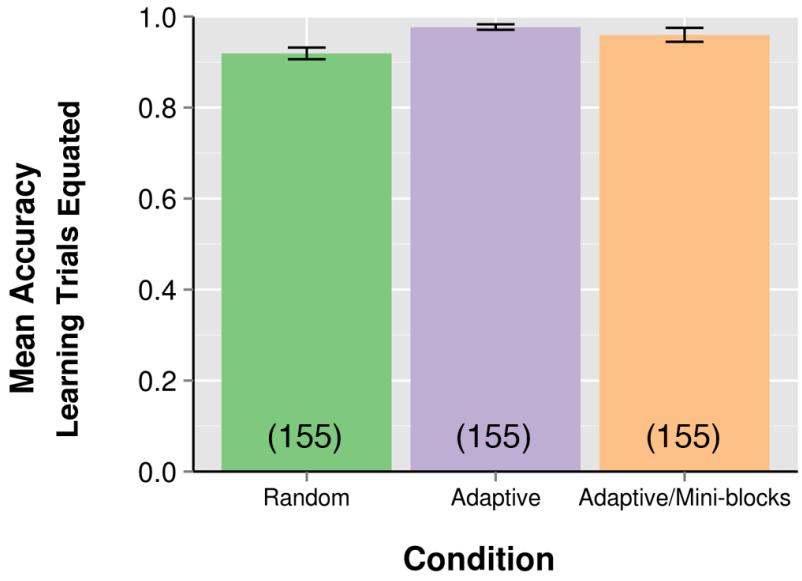

Figure 4.6: Exp 1a. Mean accuracy by learning condition based on equal numbers of learning trials. Parentheses indicate trial number at which accuracy was measured, for the two most recent presentations of each category. Error bars indicate +/- one standard error of the mean.

effect of learning condition ($F(2,51)=6.1$, $p=.004$). Individual comparisons indicated that the Adaptive condition had reliably higher accuracy than the Random condition ($p=.004$), and the Adaptive/Mini-blocks condition had marginally higher accuracy than the Random condition ($p < .06$). The two Adaptive conditions did not differ reliably ($p > .9$). (All p values were Bonferroni corrected.) There were no reliable differences in response times across conditions using a similar method of measuring RTs at an equivalent point in the three conditions (155 trials).

A final set of analyses examined mean response times (RTs) in both post-tests. Only response times from correct trials were analyzed. Response times for Experiment 1a are shown in Figure 4.7, right two panels. A 3x2x2 ANOVA examined RTs across scheduling condition, post-test phase and across novel vs. familiar stimuli. There was no reliable main

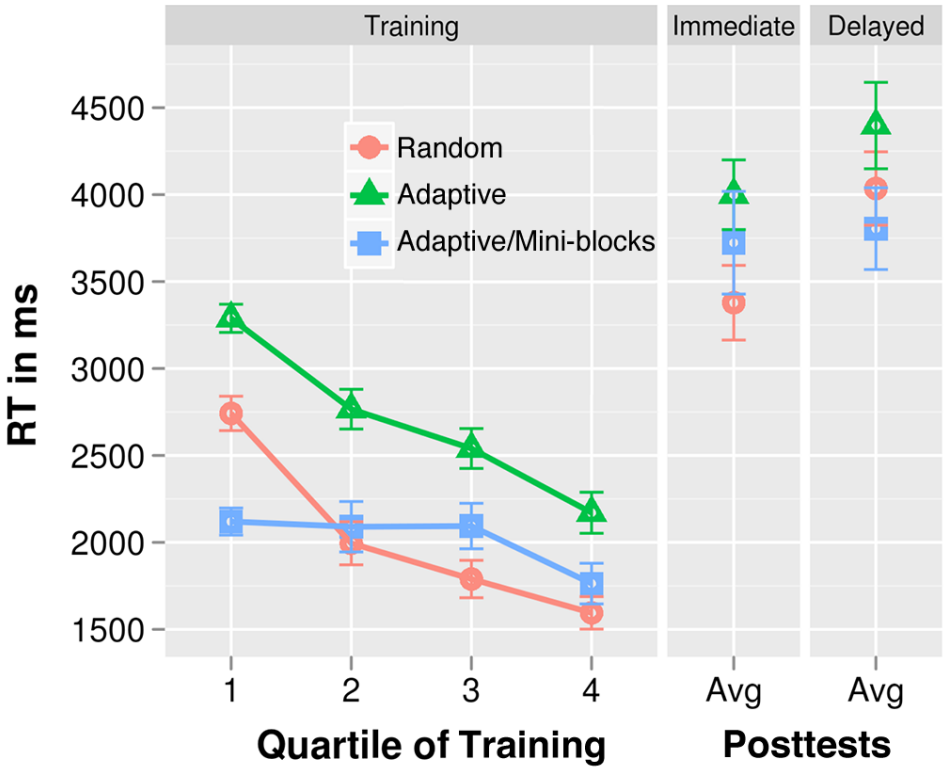

Figure 4.7: Exp 1a. Mean response times by quartile of training phase and in the immediate and delayed post-tests by scheduling condition. Response times include accurate responses only. Error bars indicate +/- one standard error of the mean.

effect of scheduling condition ($F(2,51)=1.78$, $p=.18$, $\eta^2=0.07$), a significant main effect of phase ($F(1,51)=6.54$, $p=.013$, $\eta^2=0.11$), and no effect of familiarity ($F(1,51)=2.98$, $p=.09$, $\eta^2=0.06$). There were no reliable interactions (all ps $>$ .20). RTs generally increased between immediate and delayed tests, but individual comparisons did not reveal reliable differences across post-tests for any condition (all ps $>$ .09).

4.3 Experiment 1b: Perceptual Category Sequencing of Low Variability Categories

The purpose of Experiment 1b was to investigate the effect of within-category variability on adaptive spacing in PL. The spacing principles we tested here in PL were derived from memory research using fixed items that reappear at varying intervals. In applying these concepts to PL of categories, the category, not a fixed learning item, is the unit of spacing. When it is time for another learning trial, a new instance of the category, not a repeat of an item, is presented. Intuitively, it seems that the applicability of spacing principles derived from item learning research might be greater in PL for categories with lower variability, because new instances of a given category will tend to resemble earlier ones. Recurrence of a category containing low-variability instances more closely resembles re-presentation of an identical item. This idea also applies to the ARTS adaptive learning system used here. Recall that ARTS uses response times from earlier trials to estimate learning strength. If a new exemplar of a category bears little resemblance to an earlier one, the estimate of learning strength derived for the earlier item may not predict learning strength of the current item. This problem should be more salient for high-variability categories and especially for categories that are disjunctive (i.e., an exemplar may be in the category by virtue of having either characteristic A or characteristic B). The integrity of the concept of learning strength seems likely to be greatest when it applies to an identical item recurring (as in item learning) and better for categories whose exemplars resemble each other than for those with highly variable exemplars.

4.3.1 Method

Procedure Experiment 1b replicated the procedure of Experiment 1a, but tested whether differences across learning conditions would be affected by the reduced variability of exemplars within each category. In Experiment 1b the exemplars in each category were made less variable in the following way: each category was composed of instances from one distinct

species. In Experiment 1a, each genus (category) was comprised of 3 distinct species, with 3 exemplars chosen from each of the 3 species. In Experiment 1b, only one of the original 3 species was selected for each genus, and all 9 exemplars for the category were selected from that species, effectively reducing total category variability.

Participants 54 undergraduate psychology students participated in an hour-long experiment for course credit.

4.3.2 Results

As in Experiment 1a, pre-test accuracy across conditions did not differ from chance (M=.25, SD=0.11), and an ANOVA showed no reliable differences between conditions ($F(2,51)=0.39$, $p=.68$, $\eta^2=0.02$). Paired comparisons between conditions were also not significant (all ps > .40).

Efficiency scores are shown in Figure 4.8 for the three scheduling conditions in both immediate and delayed post-tests and across novel and familiar items. Efficiency was generally higher in Experiment 1b than in Experiment 1a, as learners required fewer trials to achieve criterion performance, especially in the Adaptive condition. Efficiencies for the Adaptive condition were higher than those in the Random and Adaptive/Mini-blocks condition for both immediate and delayed post-tests (Immediate post-test: Ms= 0.18, vs. 0.10 and 0.11; Delayed post-test: Ms=0.16, vs. 0.096 and 0.095). A 3x2x2 mixed factor ANOVA with condition as a between-subjects factor, and test phase and stimulus familiarity as within-subjects factors, confirmed significant main effects of condition ($F(2,51)=8.87$, $p < .001$, $\eta^2=0.26$), test phase ($F(1,51)=84.5$, $p < .001$, $\eta^2=0.62$), and a marginal condition by test phase interaction ($F(2,51)=2.6$, $p=.084$, $\eta^2=0.092$). Paired comparisons revealed that the differences between Adaptive and both of the other two conditions (vs. Random and Adaptive/Mini-blocks) were reliable at immediate post-test ($t(34)=3.49$, $p=.001$, $d=1.22$ for Adaptive vs. Random, and $t(34)=3.07$, $p=.004$, $d=1.09$ for Adaptive vs. Adaptive/Mini-blocks) and at delayed post-test ($t(34)=3.14$, $p=.004$, $d=1.1$, and $t(34)=3.17$, $p=.003$, $d=1.12$), whereas

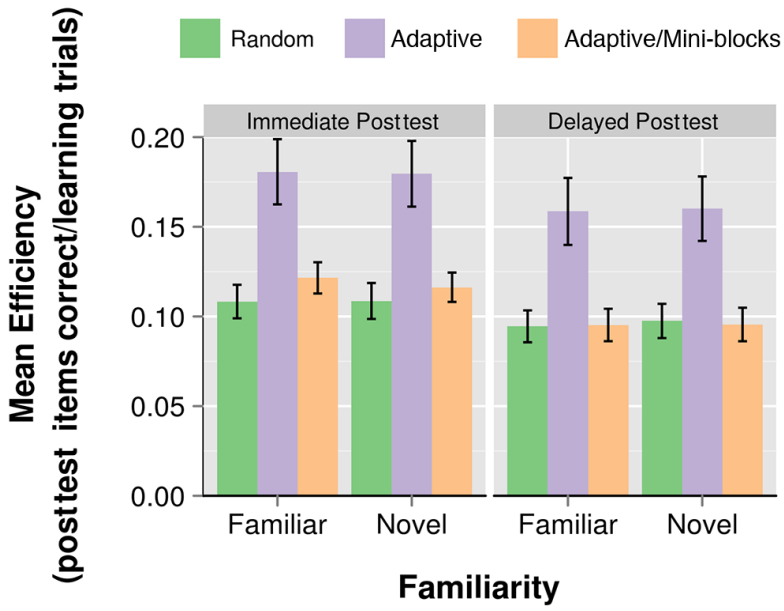

Figure 4.8: Exp 1b. Mean efficiency scores by learning condition and post-test phase. Efficiency scores were the number of post-test items answered correctly divided by the number of trials invested in learning. Familiar stimuli were post-test items that had been shown during training, whereas novel stimuli were items that had not been presented previously. Error bars indicate +/- one standard error of the mean.

the difference between Random and Adaptive/Mini-blocks conditions was not reliable (in either immediate or delayed post-test, both ps >.40). In percentage terms the Adaptive condition was 38% more efficient than Random in the immediate post-test and 39% more efficient than Random in the delayed post-test. Effect sizes for these comparisons exceeded 1.0 in both post-tests. All conditions showed a reliable decrease in efficiency between immediate and delayed post-tests (all ps $<$ .002). There was no reliable main effect of familiarity ($F(1,51)=0.01$, $p=.91$, $\eta^2<0.001$), nor any reliable interaction between familiarity and phase ($F(2,51)=1.06$, $p=.31$, $\eta^2=0.02$), familiarity and condition ($F(2,51)=0.49$, $p=.62$, $\eta^2=0.02$), or between condition and phase ($F(2,51)=0.07$, $p=.93$, $\eta^2=0.002$). The lack of main effects or interactions involving familiarity indicate that, unlike Experiment 1a, there was no

advantage in the post-tests for novel vs. previously exposed stimuli.

Trials to retirement differed between conditions; participants averaged 125.3, 174.8, and 234.9 trials in the Adaptive, Adaptive/Mini-blocks, and Random conditions, respectively. A one-way ANOVA with condition as the factor showed reliable difference ($F(2,51)=10.04$, $p < .001$, $\eta^2=0.28$). Paired comparisons indicated that all three conditions differed from one another. The Adaptive condition required fewer trials than the Random condition ($t(34) = 3.82$, $p=.001$, $d=1.36$); the Adaptive/Mini-blocks condition required fewer trials than Random ($t(34)=2.17$, $p=.037$), and the Adaptive condition required fewer trials than the Adaptive/Mini-blocks condition ($t(35)=-3.46$, $p=.002$).

Raw accuracy data (not corrected for number of trials invested) are shown in Figure 4.9 for each condition in both immediate and delayed post-tests. A 3x2 ANOVA, with scheduling condition as a between-subjects factor and post test phases as a within-subjects factor showed a marginally reliable main effect of condition ($F(2,51)=2.71$, $p=.076$, $\eta^2=0.09$), a reliable effect of test phase ($F(1,51)=86.3$, $p < .001$, $\eta^2=0.63$), and a reliable test phase by condition interaction ($F(2,51)=3.62$, $p=.034$, $\eta^2=0.12$). Accuracies in the Random condition numerically exceeded those in the Adaptive and Adaptive/Mini-blocks conditions in both the immediate and delayed post-tests (Immediate: M: .90 vs. .81 & .82, respectively; Delayed: M: .79 vs. .72 & .65, respectively). Individual comparisons confirmed a reliable difference between the Random and Adaptive conditions on the immediate post-test ($t(34)=2.19$, $p=.04$, $d=0.73$), but consistent with the observed interaction, the difference was not reliable at delayed post-test ($t(34)=1.47$, $p=.15$, $d=0.49$). No reliable differences were found in immediate post-test accuracy between the Random and Adaptive/Mini-blocks condition or between the two Adaptive conditions ($ts(34)=1.62$ and 0.29, $ps=0.11$ and 0.77 respectively). At delayed post-test, there was no reliable difference between Adaptive and Adaptive/Mini-blocks ($t(34)=1.09$, $p=.28$, $d=0.37$), but there was a reliable difference between Random and Adaptive/Mini-blocks ($t(34)=2.49$, $p=.01$, $d=0.84$).

As in Experiment 1a, we compared accuracies across conditions at a moment in training when each participant had accumulated about the same number of learning trials. Mean

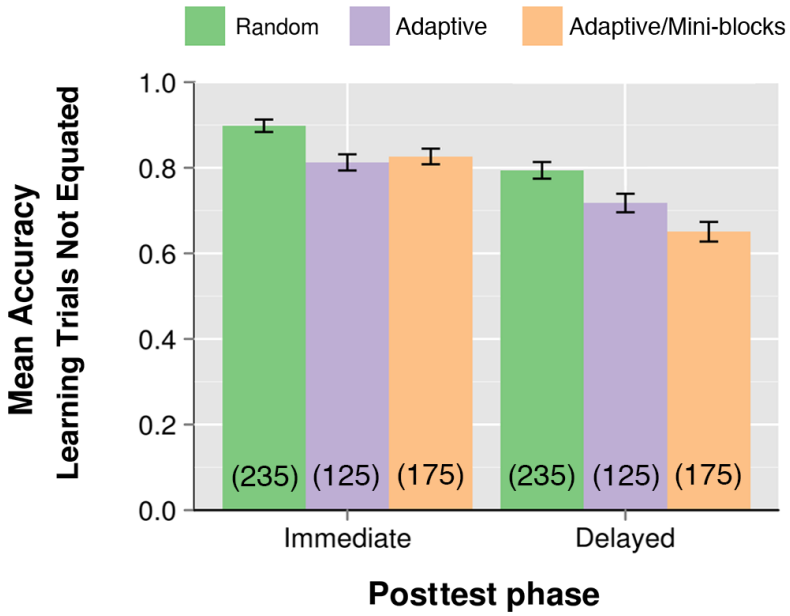

Figure 4.9: Exp 1b. Mean accuracy results by learning condition and post-test phase. Accuracy is given as the percentage of 24 post-test questions answered correctly. These data indicate raw accuracy not corrected for number of learning trials; the number of learning trials in each condition is shown in parentheses. Error bars indicate +/- one standard error of the mean.

trials to criterion was 125 in the Adaptive condition (SD=49.9), and proportion correct for the last two presentations of each stimulus category for this condition at this point in training was .99 (see Figure 4.10). In the Random and Adaptive/Mini-blocks conditions, performance measured from the 125th trial on the last two presentations of each category was .91 and .93 respectively. A one-way ANOVA comparing the learning conditions on this measure showed a reliable main effect of condition ($F(51)=7.28$, $p=.002$). Individual comparisons indicated that accuracy was reliably higher in the Adaptive condition than in the Random condition ($p=.002$) and also reliably higher in the Adaptive condition than in the Adaptive/Mini-blocks condition ($p=.035$). There was no reliable difference in accuracy after 125 trials between the Adaptive/Mini-blocks and Random conditions ($p > .83$) (all ps

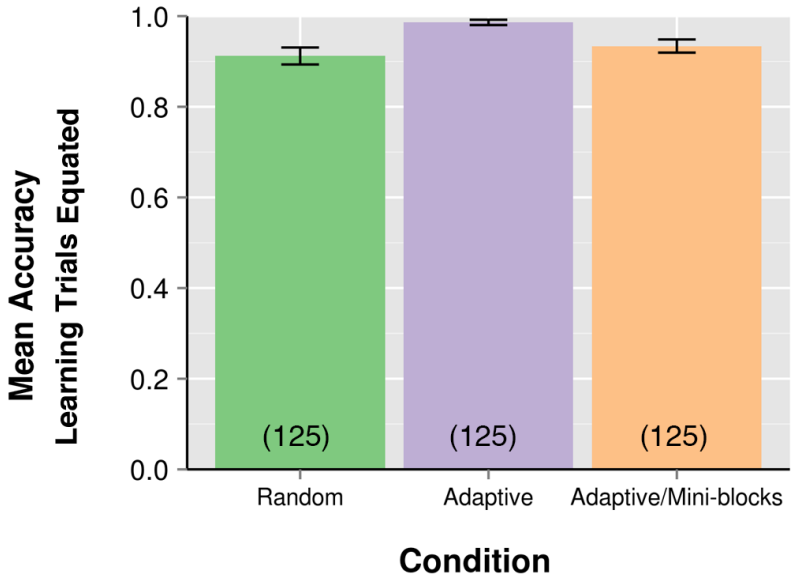

Figure 4.10: Exp 1b. Mean accuracy by learning condition based on equal numbers of learning trials. Parentheses indicate trial number at which accuracy was measured, for the two most recent presentations of each category. Error bars indicate +/- one standard error of the mean.

Bonferroni corrected).

Response times in the immediate post-test averaged 3.41, 3.48, and 2.64 sec per trial in the Adaptive, Adaptive/Mini-blocks, and Random conditions, respectively. In the delayed post-tests, response times were more similar across conditions, with the Adaptive, Adaptive/Mini-blocks, and Random conditions averaging 3.73, 3.68, and 3.33 sec per trial respectively. These observations were confirmed by a 3x2x2 ANOVA across scheduling condition, post-test phase and familiarity which found a marginally significant main effect of condition ($F(2,51)=2.83$, $p=.07$, $\eta^2=0.10$), a main effect of phase ($F(1,51)=11$, $p=.002$, $\eta^2=0.17$), and a main effect of familiarity ($F(1,51)=4.84$, $p=.032$, $\eta^2=0.86$). There were no interactions between the factors (all ps $> .20$). Examining the effect of scheduling condition, there were significantly lower RTs for the Random condition than Adaptive ($t(34)=2.37$, $p=.02$, $d=0.79$) and Adaptive/Mini-

blocks (t(34)=2.01, p=.05, d=0.69), but no difference between the two Adaptive conditions (t(34)=0.03, p=.98, d=0.01). Examining differences between immediate and delayed post-tests, there were significant increases in RT for Random ($p < .001$), but not for Adaptive or Adaptive/Mini-blocks (both $ps > .20$). Individual comparisons showed that in the immediate post-test, response times in the Random condition were shorter than in either of the other conditions (Random vs. Adaptive, t(34)=2.74, p=.02, Bonferroni corrected; Random vs. Adaptive/Mini-blocks, t(34)=3.08, $p < .012$, Bonferroni corrected). Response times did not differ between the two Adaptive conditions (t(34)=0.113, p=.91). In the delayed post-test, there were no reliable response time differences between any two conditions (all $ts(34) < 1.0$, $ps > .59$, Bonferroni corrected).

4.3.3 Efficiency and Transfer Across Experiments

We compared learning results across Experiments 1a and 1b. First, a 3x2x2x2 ANOVA with between-subjects factors of scheduling condition and experiment, and within-subject factors of post-test phase and stimulus familiarity, showed a reliable main effect of condition (F(2,102)= 10.79, $p < .001$, η^2=0.174), a large main effect of test phase (F(1,102)= 132.6, $p < .001$, 0.56), no main effect of experiment (F(1,102)= 2.30, p=.13, η^2=0.022) and a marginally significant interaction of scheduling condition and experiment (F(2,102)= 2.89, p=.06, η^2=0.053). There was also a main effect of stimulus familiarity (F(1,102)= 7.67, p=.007, η^2=0.07), and an interaction between stimulus familiarity and experiment (F(1,102)= 8.65, p=.004, η^2=0.078). No other interactions were significant (all $ps > .17$). Individual comparisons showed that the source of the main effect of condition was greater efficiency in the Adaptive condition than in either of the other conditions (Adaptive vs. Random, t(70)=3.82, $p < .001$, d=.95; Adaptive vs. Adaptive/Mini-blocks, t(70)=3.10, $p < .002$, d=.77). The Adaptive/Mini-blocks and Random conditions did not differ reliably (t(70)=1.12, p=.27, d=0.26). The same pattern of results appeared when looking separately at results from the immediate post-test (Adaptive vs. Random, t(70)=3.89, $p < .001$, d=0.96; Adaptive vs. Adaptive/Mini-blocks, t(70)=2.95, $p < .004$, d=0.73; Adaptive/Mini-

blocks vs. Random (t(70)=1.51, p=.14, d=0.36) or delayed post-test (Adaptive vs. Random, t(70)=3.64, p < .001, d=0.91; Adaptive vs. Adaptive/Mini-blocks, t(70)=3.18, p < .002, d=0.79; Adaptive/Mini-blocks vs. Random (t(70)=.67, p=.51, d=0.16).

The condition by experiment interaction was due primarily to somewhat better efficiency shown by the Adaptive condition in Exp. 1b compared to Exp. 1a (t(34)=2.03, p=.05, d=0.69). Neither the Adaptive/Mini-blocks nor Random conditions differed reliably between Experiments 1a and 1b (ps=.64 and .83 respectively). The interaction of stimulus familiarity and experiment reflects the lack of any post-test advantage for novel stimuli in Experiment 1b, unlike Experiment 1a, which showed a clear difference.

4.4 Discussion

We studied PL in a rich, natural domain that was unfamiliar to the participants. As in many real-world PL tasks, the goal of learning is to discover and encode features and relations that determine natural categories, allowing the learner to accurately classify previously unobserved instances. Specifically, we tested whether an adaptive sequencing algorithm implementing principles of spacing in an individualized manner could improve PL for natural categories. The algorithm varied intervals between presentations of new instances of each learning category based on each learner's accuracy and RT in classifying instances of that category.

Effects of adaptive sequencing on PL In both experiments, we found evidence of greater learning efficiency for adaptively sequenced learning over random presentation, in both immediate and delayed post-tests. We included a test after a one-week delay, because immediate and delayed tests sometimes differ in interesting ways, and testing after a delay removes possible influences of relatively transient effects and is therefore considered a better measure of learning (Schmidt & Bjork, 1992). In Experiment 1a, with higher variability categories (exemplars chosen from within any species in a genus), the efficiency advantage

of adaptive sequencing was clearest for novel items in the post-tests and for all items in the delayed post-test, which showed a 29% efficiency advantage over random presentation. Moreover, effect sizes for Adaptive vs. Random for both immediate and delayed post-tests averaged around .7. Adaptive sequencing also reliably outperformed random presentation on a pure accuracy measure when learning conditions were compared after the same number of learning trials. These learning effects were magnified in Experiment 1b, in which lower variability categories (using only one species per genus) were used. In this experiment, the Adaptive condition showed highly reliable advantages over the Random condition in efficiency on both immediate and delayed post-tests (on the order of 38-39%); for both familiar and novel items; and also when accuracy was compared directly across groups after the same number of learning trials. The results have significance for understanding high-level PL in general and for applications of PL in real-world education and training domains. The spacing effect is one of the most important and robust principles of learning and memory (Dempster, 1996), and with memory for factual material, adaptive learning schemes have been shown to enhance efficiency by tailoring spacing to the individual learner's course of acquisition of each item to be learned (Atkinson, 1972; Mettler, Massey, & Kellman, 2011; Pavlik & Anderson, 2008). The present work may be the first to apply adaptive spacing to PL. The present results indicate that adaptive sequencing can robustly improve learning. The effect sizes (ranging from around .7 in Exp. 1a to 1.2 in Exp. 1b), as well as the percentage advantages in efficiency (25-29% in Exp. 1a to around 38-39% in Exp. 1b) are of sufficient magnitude to be of substantial value in improving learning in complex learning domains in real-world settings.

Spacing in PL and fact learning These results may also offer some insight into relations between PL and factual learning, where spacing has been more extensively investigated. In our studies, a fundamental principle in the adaptive condition was the stretching of the recurrence interval for categories based on speed of responding. The present findings that adaptive spacing improves PL for natural categories parallel similar effects of adaptive spac-

ing for memory items. As such, it raises the question of what learning mechanisms may be shared across these domains. Storage of items in memory (fact or item learning) and discovering structure in displays that allows classification of new instances (PL) appear to involve substantially different mechanisms. In memory research, retesting an item when it is just about to be forgotten is usually considered in terms of memory trace decay (Pyc & Rawson, 2009), but in PL, learning progresses by more selective and fluent information extraction from presented displays. We believe that the common link is not that factual memory and PL involve the same mechanisms, but that a common principle of optimal learning applies to both. As learning to extract relevant information improves for one category, it becomes desirable to have a longer interval and/or more trials with intervening categories before returning for further practice on the initial category. PL involves discovery of invariance and allowable variation with categories (Gibson, 1969), but perhaps the most crucial component of PL is coming to encode distinguishing features between categories (Gibson, 1969). This process may be optimized by modulating the numbers of trials of intervening categories depending on the strength of that category. If the learner is a poor classifier of instances of a category, many intervening trials of other categories may impede learning, but as learning improves, more intervening category experiences may be optimal. Although the underlying processes for item memory and PL are unlikely to be the same, learning in both domains can be enhanced by adjusting spacing to match changes in learning strength. And in both domains, because learning strength may not be predictable in advance and may vary by learners and categories, adaptive scheduling based on updated learning strength estimates, as was done here by the use of response times, may offer advantages over predetermined schedules.

Category variability in PL The advantage for lower variability categories can be easily interpreted in this context. The ARTS system uses response times, along with accuracy, from earlier trials to estimate learning strength. When a specific item recurs, as in factual learning contexts, an accurate and faster response can be straightforwardly interpreted as

an improvement in learning. A primary goal of the present work was to investigate if the accuracy and speed can also be used to guide PL, where categories recur over spacing intervals but presented instances are novel. The results indicate that adaptive sequencing of categories is indeed beneficial, but the benefit is greater for categories with lower variability among instances. The instances of higher variability categories may involve a greater array of features and relations to be encoded; thus, a learner's performance on an earlier instance of a category may be an imperfect predictor of learning strength for another item. High variability categories might even be disjunctive, in the sense that there is more than one characteristic that confers membership, or in the sense that irrelevant variation may differ from instance to instance. Where such differences exist, performance measured from one instance might provide little indication of the learning strength for another instance. Gibson's classic work on PL emphasized discovery of invariance, but many natural categories may have a family resemblance structure (Rosch & Mervis, 1975; Wittgenstein, 1953). Perhaps even more crucially, the process of discovering distinguishing features of categories (Gibson, 1969) may also involve learning to ignore characteristics that are not diagnostic of category membership. These may also vary across instances of a single category.

We close this issue by noting that the situation may actually be more complicated. PL in category learning involves the discovery and selective encoding of diagnostic characteristics that govern category membership. Explanations of PL based on selection have been supported by considerable empirical and modeling work (e.g., Petrov, Dosher, & Lu, 2005). In PL contexts involving categorization of complex, multidimensional stimuli, one implication of selection is that, as learning progresses, members of the same category will likely come to resemble each other more. In this sense, the perceived 'variability' of instances of a category likely changes through PL. It might be interesting in future research to develop measures of perceived similarity to look at stimulus variability as a dependent variable that changes in PL, in addition to its effects as an independent variable as in the present research.

Transfer A hallmark of perceptual learning in real-world domains is transfer of learning. Learners become able to accurately and fluently classify new exemplars of previously learned categories. To ascertain that true PL, rather than memorization of instances, was involved in the present studies, we used post-tests with both familiar and novel instances. All of our results indicate that novel instances were classified at least as accurately as familiar instances. These outcomes indicate both that participants attained classification skills that generalized to previously unseen cases and also that our efforts to minimize instance repetitions during learning were successful.

The results of Exp. 1a actually suggested better performance for novel exemplars, and this tendency was strongest in the Adaptive condition at delayed post-test. While we cannot rule out some possible effect of interest here, this seems to us to be most likely an inconsequential finding. The set of novel exemplars used in the post-tests was the same for all conditions, and this set may have simply been, on average, slightly less difficult than the familiar instances used in the post-test. No advantage for novel items appeared in Experiment 1b, which used a different, fixed set of post-test items. If, paradoxically, there is some reason that PL in some conditions is actually stronger for novel instances, the current experiments were not designed to reveal this clearly. Use of a "jackknife" procedure, where each subject is presented different novel instances, would be preferable for a study focused on this issue. Our use of novel and familiar post-test items allowed clear comparisons across conditions, and provided clear evidence for transfer of learning, but it did not provide clear evidence for a novelty advantage.

Partial blocking in PL Our data offer little or no support for initial blocking or massing of instances of a given category. At best, the Adaptive/Mini-blocks condition in the present experiments produced performance nearly equivalent to the Adaptive condition; it was often somewhat worse, and never better. The intuition behind blocking is that learning of commonalities within each category should be facilitated by seeing several instances in succession. This intuition appears to be incorrect, however. Earlier work (Kornell & Bjork, 2008) com-

pared complete blocking to complete interleaving in studying examples of different artists' painting styles and found a clear advantage of interleaving. In our Adaptive/Mini-blocks condition, we investigated whether some initial blocking, followed by interleaved, spaced practice might aid early learning of categories while still capturing the benefits of interleaving later. This approach never produced better performance than the regular Adaptive condition, in which there was consistent interleaving. It appears that stimulus presentation that facilitates the learning of contrasts that distinguish categories may be of greatest importance in arranging PL.

Learning to criterion in PL The studies reported used learning to criterion. Probably for reasons of experimental control, this is quite rare for studies of spacing in learning. It is, however, of primary importance in real learning settings. The most obvious methodological difficulty of studies using learning to criterion is that different participants and conditions will require different numbers of learning trials. The efficiency measure addressed this issue by combining both post-test accuracy and the number of learning trials invested; such a measure is likely to be useful in real-world learning settings where mastery in the shortest time is desirable. As reaching criterion in our Random condition generally required more trials than in the Adaptive condition, it is important to consider whether this feature alone provided the advantage of adaptive sequencing. To address this issue, we also examined accuracy after a similar number of learning trials in each group. As with the efficiency measure, this "apples to apples" comparison of accuracy also clearly showed advantages of adaptive sequencing.

A final note concerns our choice of comparison conditions. We chose to compare sequencing algorithms against random presentation - a notoriously effective schedule of practice that produces robust, albeit inefficient, learning in a variety of contexts. For example, random presentation automatically implements a type of spaced interleaving, and when unmodified, as in our experiment, can result in repeated presentation of critical stimulus material. Though random presentation has fared poorly in some experiments that have compared scheduling

algorithms to random practice (as in Atkinson, 1972), it may be that in a learning domain with as few categories as in our experiment (12 categories), the benefits of random presentation may be quite large (compared to, for instance, learning a large number of independent factual items). The fact that our algorithms performed as well as they did is thus encouraging. Presumably, if we had compared adaptive sequencing to massed practice (blocking of all category exemplars) adaptive sequencing would have fared even better (c.f., Kornell & Bjork, 2008).

4.5 Experiment 2: Perceptual Category Learning on Fluency with 3D Chemical Representations

4.5.1 Introduction

As discussed in Chapter 3, learning to criteria is an important learning feature that interacts with the size of spacing delays and difficulty of retrieval during criterial practice. While somewhat under-studied on its own, learning criteria that also include response speed are rare in both the experimental literature as well as in practical applications of learning criteria to learning. In order to further investigate the relationship between response speed and learning criteria in adaptively spaced practice, we conducted an experiment that manipulated the degree of reliance on response speed in learning criteria. The experiment was designed to test the effects of response speed dependent criteria on perceptual learning.

Using software-based, web-delivered, perceptual learning modules (PLMs), students in introductory community college chemistry classes gained knowledge of various chemical structures including discrimination of bond angles and determination of hybridization. PLMs presented students with up to 3 related representations of a chemical structure: animated 3D (3-dimensional) representations, 2D Lewis diagrams, and chemical formulas. PLMs utilized an adaptive sequencing algorithm that monitored students accuracies and response times to determine the trials that should elapse between repeated presentations of structural

categories. Students procceded through stages that progressively decreased the number of simultaneously available structural representations until they were able to answer questions about chemical structure solely with reference to chemical formulas. We tested two adaptive sequencing conditions, one that included response time and accuracy in the determination of item retirement and one that included accuracy alone. Results showed significant learning gains in both conditions, and an effect of response-time-based learning criteria on students' long-term gains in fluency (speed) at making determinations of chemical structure.

4.5.2 Methods

Participants The participants were 28 community college students who were completing the experiment as a component of an introductory chemistry course (GenChem 1 at Collin Community College, Collin, Texas). One participant did not complete a delayed post-test.

Design There were two between subjects conditions that presented learning items using an adaptive scheduling algorithm but differed in whether the learning criteria utilized response speed (RT condition) or not (no RT condition). Response criteria in the RT condition were set at 12 seconds. The experiment utilized a pre-test/post-test design with a 2 month delayed post-test. There were three similar versions of the test assessment, and participants received a different version at pre-test, post-test and delayed post-test.

Procedure Participants were randomly assigned to either the RT or noRT condition. Participants completed a number of sessions. At the first session participants completed a pre-test, and began the learning session. Students continued working on the module over the next few days, both in class and at home, until reaching mastery for all items. Seven days after the first session they completed a post-test. After two weeks, participants took a delayed post-test in class. Assignment of test versions to pre, post and delayed post-test were counterbalanced across participants.

During learning sessions, participants responded to multiple choice questions in a PLM.

Initially participants were shown an animated 3D diagram, a Lewis structure and a bare chemical formula (see Figure 4.11) and were asked various questions about either bond angle or hybridization of the displayed structure. Feedback was presented on each trial. Over the course of learning, representations were removed in a fading process: first 3D diagrams, then Lewis structures were removed, until, at the end of learning, students were answering questions based on the bare formulas alone. In order to progress between sections, students had to meet learning criteria of four consecutive correct responses per category. In the RT condition, students had to additionally respond faster than 12 seconds at each criterial trial.

Materials Categories of problems consisted of bond angle identification problems of four types (180, 120, 109.5, 90 degrees) as well as hybridization problems of 6 types (s, sp, sp2, sp3, sp3d, sp3d2). See Figure 4.11 for some examples of trial types. Additionally, categories consisted of 3 levels of difficulty - a) easy trials, where a 3D diagram, a Lewis structure, and a formula were shown, b) medium trials, where the 3D diagram was removed and only a Lewis structure and formula were shown, and c) difficult trials, where only a formula was shown. All learning session trials were followed by feedback for the correct answer after the learner's response. Chemical structures were chosen from a set of common compounds typical of the level of the chemistry students in the course. There was also input from the instructor on the choice of chemical structures prior to the start of the experiment.

Pre-test and post-test materials differed from learning session trials. Pre-test and post-test trials required multiple choice responses, and did not present feedback after each response. Trials usually showed a single form of representation, e.g., a single Lewis diagram, and asked to choose the correct bond angle or hybridization from a multiple choice list. There were also problems that asked for the reverse mapping, showing a hybridization value and asking for the Lewis structure or 3D diagram from a choice of three diagrams.

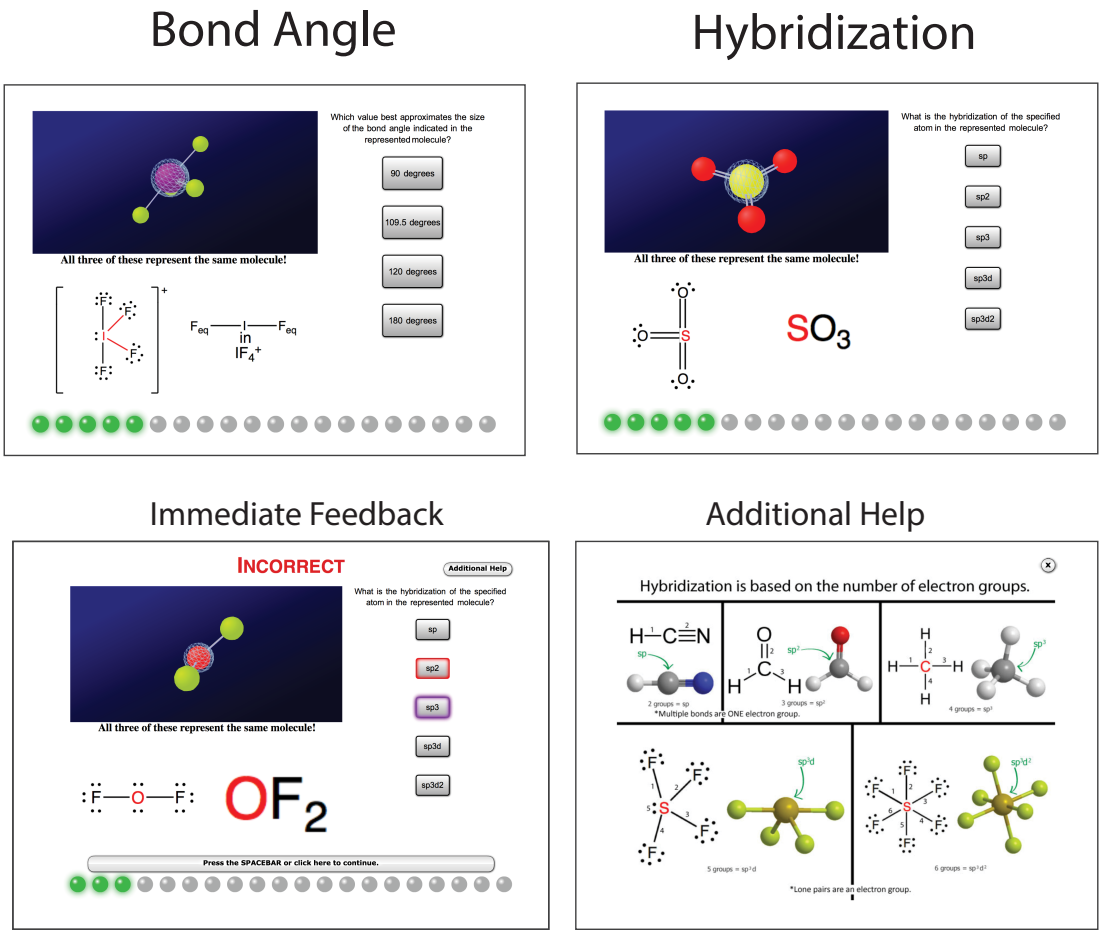

Figure 4.11: Exp 2. Example of trial screens showing bond angle and hybridization problems. Immediate feedback was presented after every response. Additional help could be triggered with a key press during the feedback phases.

Planned analyses The primary analyses in Experiment 2 were accuracy change scores and latency change scores. It was predicted that an adaptive algorithm that monitors learning strength for category learning will result in greater learning, measured by both accuracy increases as well as response speed decreases (fluency), when learning criteria take into account response speed (RT condition) than when response speed is not a component of

learning criteria (no RT condition).

4.5.3 Results

Raw accuracies Accuracies are shown in Figure 4.12. A 2x2 mixed factor ANOVA on condition (RT vs. noRT) and test phase (post-test vs. delayed post-test), found no effect of condition ($F(1,25)=2.24$, $p=0.15$), a significant main effect of test phase ($F(1,25)=13.31$, $p<.01$), and no condition by test phase interaction ($F(1,25)=0.24$, $p=.63$). Accuracies declined between post-test and delayed post-test ($ps<.05$).

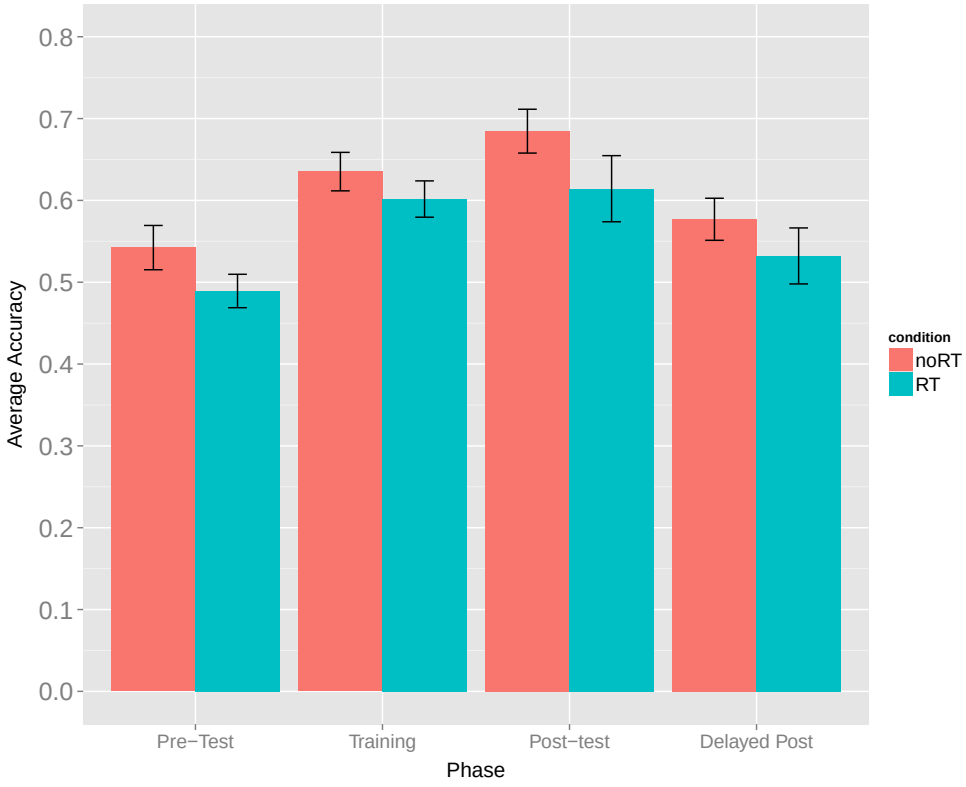

Figure 4.12: Exp 2. Accuracy across experiment phase and by mastery condition (RT vs noRT). Error bars show +/- 1 standard error of the mean.

Accuracy change score Accuracy change scores (accuracy on post-test minus accuracy on pre-test) are shown in Figure 4.13. A 2x2 mixed factor ANOVA on condition and

test phase found no effect of condition (F(1,25)=0.01, p=.93), a main effect of test phase (F(1,25)=13.31, p=.001), and no condition by test phase interaction (F(1,25)=0.24, p=.62). Accuracy change scores declined between post-test and delayed post-test for the noRT condition (t(12)=4.78, p<.001), but not reliably for the RT condition (t(13)=1.81, p=.09).

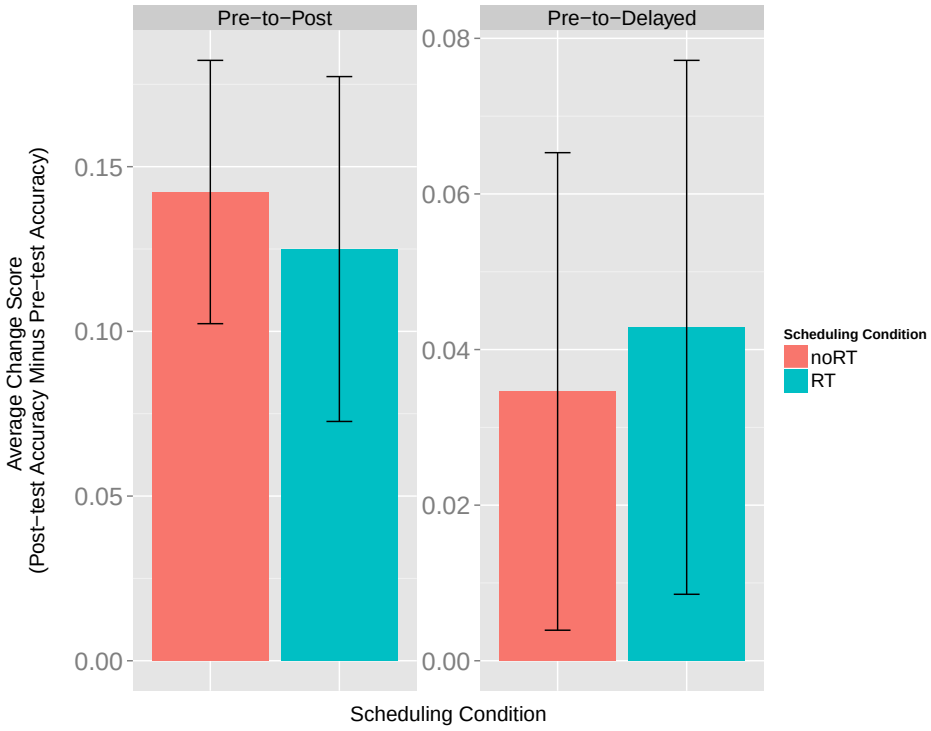

Figure 4.13: Exp 2. Accuracy change scores (post-test accuracies minus pre-test accuracies) by post-test phase (initial post-test vs. 2 month delayed post-test) and mastery condition. Error bars show +/- 1 standard error of the mean.

Response time There was an effect of response-time-based learning criteria on students' long-term gains in fluency - or speed of responding at making determinations of chemical structure. Fluency was measured as a decrease in response time for responses at post-test compared to at pre-test (it was possible to make this comparison due to the high levels of prior knowledge - accurate responses - at pre-test). Response-time reductions are shown in

Figure 4.14. A 2x2 mixed factor ANOVA on condition and test phase found no main effect

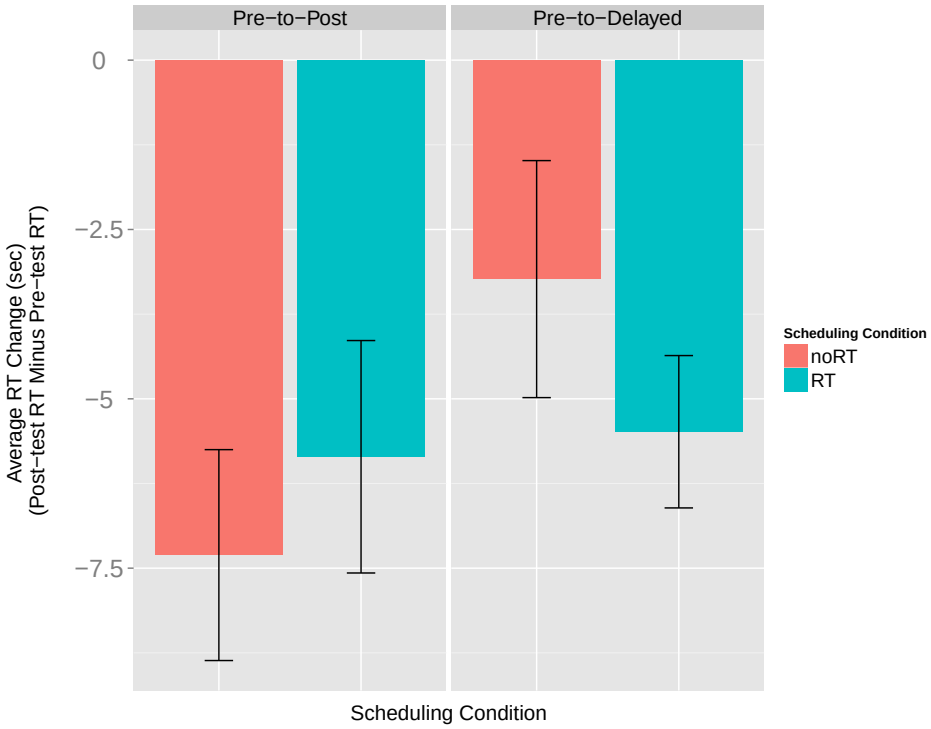

Figure 4.14: Exp 2. Response-time change scores (post-test RTs minus pre-test RTs) by post-test phase (initial post-test vs. 2 month delayed post-test) and mastery condition. RTs are from correct responses only. Error bars show +/- 1 standard error of the mean.

of condition ($F(1,25)=0.04$, $p=.84$), a main effect of test phase ($F(1,25)=4.65$, $p=.04$), and a marginally significant condition by test phase interaction ($F(1,25)=3.44$, $p=.075$). T-tests between each condition were not significant at either test (post-test: $t(25)=0.623$, $p=.54$; delayed post-test: $t(25)=1.1$, $p=.28$), and the paired t-tests showed that the difference between the post-test and delayed post-test for the noRT condition was significant ($t(12)=1.74$, $p=.003$), but the difference between the post-test and delayed post-test for the RT condition was not significant ($t(13)=0.23$, $p=.82$). This interaction indicated that the no-RT condition showed greater fluency gains at an immediate post-test (M=7.3 sec reduction, vs. 5.8 sec reduction in the RT condition, SDs=5.6,6.4 respectively), but the RT condition had greater

fluency gains at a delayed test, two months later (RT: M=5.4 sec, SD=4.2 vs. noRT:3.2 sec, SD=6.3). Those in the RT-based criteria condition appeared to maintain their fluency gains after a delay of 2 months, while students in the no-RT condition lost a significant portion of their gains.

4.5.4 Discussion

It appears that when learning complex structural information in a chemistry course, the use of a PLM (perceptual learning module) increased students' fluency with determining aspects of chemical structure. Across two mastery criteria conditions, significant gains in fluency were found between pre-test and post-test. In addition, these effects lasted for over two months, as measured by gains on a delayed post-test. No significant differences were found between scheduling conditions in terms of the amount learned, although the sample size was particularly small, and participant compliance (e.g., studying at home) was particularly uncontrolled. Nevertheless, the study gives us confidence that adaptive scheduling techniques, combined with stringent mastery criteria can provide fluency advantages to real-world coursework and in realistic timescales, during the typical schedule of a chemistry course. The study demonstrates that many of the practical and theoretical points about scheduling, learning criteria and fluency are applicable to materials that are richly organized, perceptual, and relevant to real learning: where learning consists of categories of stimuli, where there are relations between stimuli within and between categories, and where the material is more than a collection of facts.

4.6 Conclusion and Discussion of Adaptive Sequencing in Perceptual Category Learning

4.6.1 Conclusion of Experiments 1a, 1b, and Experiment 2

It is becoming increasingly clear that perceptual learning comprises a pivotal component in domains where humans attain high levels of expertise, including high-level cognitive domains that have traditionally been considered to have little to do with perception (for recent reviews, see Kellman & Garrigan, 2009; Kellman & Massey, 2013). More than one aspect of perceptual learning is important, including both discovery effects – finding the information relevant to a classification – and fluency effects – coming to handle the input quickly and/or with lower cognitive load (Gibson, 1969; Goldstone, 1998; Kellman & Garrigan, 2009; Shiffrin & Schneider, 1977). Perhaps most important in complex tasks is discovery of structural information amidst task-irrelevant variation (Gibson, 1969), with the hallmark of this kind of PL being that the learner can accurately and fluently classify previously unseen instances. Whether we consider a child who learns to see a new animal and correctly say "cat," the skilled instructor who accurately derives language structure from a student's poor handwriting, the "chick sexers" described by Gibson (1969) or the scientist intuitively grasping patterns in equations and graphs, the discovery of relevant structure and the ability to use it in new cases is important.

Both basic research and understanding of the widespread implications of perceptual learning raise questions about how to optimize it. Although a great deal of work has been done to understand principles of factual or procedural learning, relatively little work has asked these same questions about PL. No previous studies that we know of have investigated how adaptive spacing techniques might fare when learning consists, not of the memorization of words or facts, but in attuning perceptual systems to extract structure. Here we have shown that adaptive scheduling strategies that enhance declarative learning domains also apply robustly to learning perceptual classifications.

Experiments 1a and 1b showed that adaptive techniques lead to more efficient perceptual learning; these effects are strongest when categories have less internal variability rather than more; and the effects lead to transfer in classifying novel instances that is fully as accurate as performance on cases previously observed.

In addition, Experiment 2 showed that the effects of learning complex perceptual categories with mastery criteria that enforce fast responses are to increase and maintain the degree of fluency developed when discriminating representations. Mastery criteria that require speed and accuracy in order to proceed to more difficult levels, or in order to retire items and finish learning, build on concepts of mastery learning in the education and cognitive sciences that are the building blocks of sound educational approaches to fostering durable, long-term learning. In addition, mastery criteria appear to cooperate with the spacing advantages of adaptively scheduled learning - to enforce both learning goals as well as spacing delays, in order to promote long-term learning.

CHAPTER 5

Conclusion

5.1 Summary of Dissertation, Results and Importance of Work

5.1.1 Summary of Dissertation

We introduced an algorithm that adapts to learners, scheduling the order of learning trials dynamically – adaptive sequencing. The adaptive algorithm responds to learners' ongoing performance and attempts to generate space or delay between presentations of items in order to improve the quality and duration of learning for those items. Decisions about how to schedule trials are based on the spacing effect in memory, the finding that longer durations between repeated presentations of learning material are more beneficial than shorter ones (Glenberg, 1976; Karpicke & Bauernschmidt, 2011). The best spacing durations are those that maximize the amount of time before items are repeated but are frequent enough to prevent forgetting.

We then demonstrated, across 6 experiments, how adaptive sequencing improves learning when compared to a variety of types of schedules of practice, and in a variety of learning domains.

5.1.2 Summary of Results

Adaptive sequencing improved learning when compared to fixed schedules of practice, both in controlled experiments where the number of presentations was limited (Chapter 2, Experiments 1 and 2), and when experiments involved more realistic durations of practice (Chapter

3, Experiments 1 and 2). In the case where number of presentations were fixed, learning improvements consisted of accuracy gains. In the case where presentations continued until learning criteria were met, adaptive sequencing improved the efficiency of learning, that is, the rate of learning. In all cases comparing adaptive to fixed schedules, learning gains were demonstrated after a delay (1 week), indicating that learning gains were durable.

Adaptive sequencing was also demonstrated in a number of learning domains. Learning improvements were shown when learning consisted of memorization of basic facts in a continuous paired-associate procedure (Chapter 2 and 3: country names and locations) and when learning involved perceptual learning - pick up of information that enables discrimination of categories of stimuli (Chapter 4: categories of butterfly species in experiments 1a and 1b; discrimination of structure in chemical compounds in experiment 2). In the case of perceptual learning, learning gains were measured in terms of *efficiency* – a measure of the rate of learning combined with overall retention of information – as well as fluency, or the speed of responding. Learners who learned to discriminate taxonomically organized images of butterfly species using an adaptive category sequencing algorithm, improved their learning at a faster rate than those who had learning trials scheduled randomly. Learners who learned to discriminate 3-dimensional structural characteristics of chemical compounds using an adaptive sequencing algorithm that incorporated learning criteria based on response speed and accuracy, became faster at making classifications and maintained those speed improvements for longer durations (2 month delay) than learners whose schedules used learning criteria based only on accuracy. Such learning gains are crucial components of perceptual learning, and are key drivers of expertise in many high level human skills (Gibson, 1969; Kellman, 2002; Kellman & Garrigan, 2009).

5.1.3 Discussion

It is well known that stretching the spacing interval during learning produces long-lasting learning in many domains and across many timescales (Dempster, 1989; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006). It is thought that a likely explanation is that spacing intervals

are powerful when delays between retrievals make recall difficult but not impossible (Bjork & Bjork, 1992; Thios, D'Agostino, 1976). What is less well understood is exactly how to take advantage of spacing effects in real learning situations: how to schedule learning across multiple, spaced presentations, and how to incorporate learning criteria that combine with spacing to make learning efficient and durable. The goal is that a learner should be able to go for longer periods of time before requiring renewed learning opportunities. In this thesis we have presented a suggested solution: that measurements of response speed during learning can provide a window into the ongoing learning strength of learners - predicting spacing delays that will maximize the benefit to learning. Adaptive schedules can incorporate learning criteria that ensure content is mastered in the shortest amount of time and with the greatest benefits to long-term learning strength. Further, in an adaptive scheduling scheme, learning criteria that ensure fast responses help to guarantee long spacing delays between criterial practice, increasing the durability of learning while also decreasing the total number of practices - that is, increasing not just raw learning gains, but the efficiency of practice.

5.1.4 Importance and Future work

It is hoped that the research in this thesis will help to accelerate processes of discovery in the application of cognitive models to real-world learning, in two ways. First, it is hoped that adaptive schedules will be applied in real-world learning situations. Many of the studies in this thesis provide evidence that adaptive sequencing is a powerful tool to increase learning in a number of domains. Some of the work reported here develops links that are needed both to understand the generality of learning principles and to develop useful, real-world applications. One link is between the development of principles of memory and learning based on laboratory studies having only a few presentations to situations in which learning proceeds to some standard of mastery. Another is the link between the literature on fixed spacing of items and the literature on adaptive learning, two valuable areas of research that have rarely been connected or compared. The final link is the application and study of basic principles of learning to diverse material and types of learning, including real-world students

and content domains.

Second, it is hoped that the development and testing of adaptive scheduling can help to further research on learning and memory, especially the spacing effect. Adaptive schedules can potentially generate optimal presentation delays, in a way that could not otherwise be determined by manual exploration of possible sequencing schedules. And since, as has been discussed, matching schedules to ongoing learning strength is an important prerequisite for examining other conditions of spaced practice (e.g., varying learning criteria), it is hoped that adaptive techniques can be used as a tool to address related experimental questions about learning.

There is also a pressing need to find models of response time and accuracy criteria in learning - for instance, convolving the probabilities of response success across multiple consecutive presentations in learning with the probability distributions of possible response speeds is an important step that could help generate new insight into appropriate learning criteria for developing mastery. In addition, there may be ways of forming complex learning criteria based on global learning set strength, or relations between categorically connected learning items.

Prior successes in the laboratory and in schools give us confidence that methods of adaptive scheduling are applicable generally (Kellman, Massey & Son, 2009). We look forward to refining these methods and extending them to larger educational problems.

Appendix A

Appendix

Table A.1: Adaptive sequencing default parameters.

Parameter	**Value**
a — Counter weight	0.1
r — RT weight	1.7
W — Incorrect penalty	20
D — Enforced delay constant	2
P_d — Default priority for unpresented items	1.1

References

Anderson, J. R. & Lebiere, C. (1998). *Atomic components of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, 2(6), 396–408.

Atkinson, R. C. (1968). Computerized instruction and the learning process. *American Psychologist*, 23(4), 225–239.

Atkinson, R. C. (1972). Optimizing the learning of a second-language vocabulary. *Journal of Experimental Psychology*, 96(1), 124–129.

Benjamin, A. & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, 61(3), 228–247.

Benjamin, A. S. & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Hillsdale, NJ: Erlbaum.

Bjork, R. A. & Allen, T. W. (1970). The spacing effect: Consolidation or differential encoding? *Journal of Verbal Learning and Verbal Behavior*, 9(5), 567–572.

Bjork, R. A. & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes*, volume 2 (pp. 35–67). Hillsdale, NJ: Erlbaum.

Bloom, B. S. (1968). Learning for mastery. *Evaluation Comment*, 1(2), 1–12. (UCLA–CSEIP).

Bloom, B. S. (1974). Time and learning. *American Psychologist*, 29(9), 682–688.

Bransford, J., Brown, A., & Cocking, R. (2000). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academies Press.

Brown, J., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher*, 18(1), 32–42.

Bush, R. R. & Mosteller, F. (1955). *Stochastic models for learning.* New York: Wiley.

Callender, A. A. (2010). Expanding retrieval promotes long term retention by preventing rapid rates of forgetting. In *Proceedings of the 32rd Annual Meeting of the Cognitive Science Society*.

Carpenter, S. K. & Delosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition*, 34(2), 268–276.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition*, 16(2), 438–448.

Carvalho, P. F. & Goldstone, R. L. (2011). Sequential similarity and comparison effects in category learning. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society.* Boston, MA: Cognitive Science Society.

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., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, 19(11), 1095–1102.

Chandler, P. & Sweller, J. (1991). Cognitive load theory and the format of instruction. *Cognition and Instruction*, 8(4), 293–332.

Chomsky, N. (1975). *Reflections on language.* New York: Pantheon Books.

Ciccone, D. S. & Brelsford, J. W. (1976). Spacing repetitions in paired-associate learning: Experimenter versus subject control. *Journal of Experimental Psychology: Human Learning and Memory*, 2(4), 446.

Crowder, R. (1976). *Principles of learning and memory.* Hillsdale, NJ: Erlbaum.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, 14(3), 215–235.

Cull, W. L., Shaughnessy, J. J., & Zechmeister, E. B. (1996). Expanding understanding of the expanding-pattern-of-retrieval mnemonic: Toward confidence in applicability. *Journal of Experimental Psychology: Applied*, 2(4), 365–378.

Dempster, F. (1996). Distributing and managing the conditions of encoding and practice. *Memory*, 10, 317–344.

Duda, R. O. & Hart, P. E. (1973). *Pattern recognition and scene analysis.* New York: Wiley.

Ebbinghaus, H. (1913). *Memory: A contribution to experimental psychology.* New York: Teachers College, Columbia University.

Eich, J. M. (1982). A composite holographic associative recall model. *Psychological Review*, 89(6), 627–661.

Estes, W. K. (1950). Toward a statistical theory of learning. *Psychological Review*, 57(2), 94–107.

Gibson, E. (1969). *Principles of perceptual learning and development.* New York: Appleton Century Crofts.

Glenberg, A. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, 15(1), 1–16.

Goldstone, R. (1998). Perceptual learning. *Annual Review of Psychology*, 49(1), 585–612.

Goldstone, R. (2000). Unitization during category learning. *Journal of Experimental Psychology: Human Perception and Performance*, 26(1), 86–112.

Hanson, N. R. (1969). *Perception and discovery.* San Francisco: Freeman, Cooper.

Hebb, D. O. (1949). *The organization of behavior: A neuropsychological approach.* New York: Wiley.

Holland, J. G. & Skinner, B. F. (1961). *The analysis of behavior: A program for self-instruction.* New York: McGraw-Hill.

Hovland, C. I. (1951). Human learning and retention. In S. S. Stevens (Ed.), *Handbook of experimental psychology.* New York: Wiley.

Hull, C. L., Hovland, C. I., Ross, R. T., Hall, M., Perkins, D. T., & Fitch, F. B. (1940). *Mathematico-deductive theory of rote learning: a study in scientific methodology.* New Haven, CT: Yale University Press.

Jenkins, J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. C. . F. I. M. Craik (Ed.), *Levels of processing and human memory* (pp. 429–446). Hillsdale, NJ: Erlbaum.

Jin, I., Kandel, E. R., & Hawkins, R. D. (2011). Whereas short-term facilitation is presynaptic, intermediate-term facilitation involves both presynaptic and postsynaptic protein kinases and protein synthesis. *Learning & Memory*, 18(2), 96–102.

Jost, A. (1897). Die assoziationsfestigkeit in abhangigkeit von der verteilung der wiederholungen. *Zeitschrift fur Psychologie*, 14, 436–472.

Kahana, M. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24(1), 103–109.

Kandel, E. R. (2001). The molecular biology of memory storage: a dialogue between genes and synapses. *Science*, 294(5544), 1030–1038.

Kang, S. H. & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, 26(1), 97–103.

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), 1250–1257.

Karpicke, J. D. & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, 331(6018), 772.

Karpicke, J. D. & Roediger, 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), 704–719.

Karpicke, J. D. & Roediger, H. L. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition*, 38, 116–124.

Kellman, P. J. (2002). Perceptual learning. In H. Pashler & R. Gallistel (Eds.), *Stevens' handbook of experimental psychology, Vol. 3: Learning, motivation, and emotion* (pp. 259–299). New York, NY: John Wiley & Sons.

Kellman, P. J. & Garrigan, P. (2009). Perceptual learning and human expertise. *Physics of Life Reviews*, 6(2), 53–84.

Kellman, P. J. & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation*, volume 58 (pp. 117–165). Academic Press, Elsevier Inc.

Kellman, P. J., Massey, C. M., & Son, J. Y. (2009). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science*, 2(2), 285–305.

Kornell, N. & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, 14(2), 219–224.

Kornell, N. & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"? *Psychological Science*, 19(6), 585–592.

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, 25(2), 498–503.

Krueger, W. (1930). Further studies in overlearning. *Journal of Experimental Psychology*, 13(2), 152–163.

Kuai, S.-G., Zhang, J.-Y., Klein, S. A., Levi, D. M., & Yu, C. (2005). The essential role of stimulus temporal patterning in enabling perceptual learning. *Nature Neuroscience*, 8(11), 1497–1499.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by.* Chicago: University of Chicago Press.

Landauer, T. K. (1975). Memory without organization: Properties of a model with random storage and undirected retrieval. *Cognitive Psychology*, 7(4), 495–531.

Landauer, T. K. (2010). Distributed learning and the size of memory. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A festschrift in honor of Robert A. Bjork* (pp. 49–69). New York, NY: Psychology Press.

Landauer, T. K. & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. Gruneberg, P. Morris, & R. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Leitner, S. (1972). *So lernt man lernen*. Freiburg: Herder.

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.

Lyon, D. O. (1914). The relation of length of material to time taken for learning, and the optimum distribution of time. Part I. *Journal of Educational Psychology*, 5(1), 1.

Marr, D. (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, (pp. 161–234).

Mettler, E. & Kellman, P. J. (2009). Concrete and abstract perceptual learning without conscious awareness. *Journal of Vision*, 9(8), 871–871.

Mettler, E., Massey, C. M., & Kellman, P. J. (2011). Improving adaptive learning technology through the use of response-times. In *Proceedings of the 33rd Annual Meeting of the Cognitive Science Society*.

Mondria, J.-A. & Mondria-De Vries, S. (1994). Efficiently memorizing words with the help of word cards and "hand computer": Theory and applications. *System*, 22(1), 47–57.

Mozer, M., Pashler, H., Cepeda, N., Lindsey, R., & Vul, E. (2009). Predicting the optimal spacing of study: A multiscale context model of memory. In *Advances in Neural Information Processing Systems*, volume 22 (pp. 1321–1329).

Naqib, F., Farah, C. A., Pack, C. C., & Sossin, W. S. (2011). The rates of protein synthesis and degradation account for the differential response of neurons to spaced and massed training protocols. *PLoS computational biology*, 7(12), e1002324.

Newell, A. & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–5). Hillsdale, NJ: Erlbaum.

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 29(6), 1051–1057.

Pavlik, P. I. & Anderson, J. R. (2008). Using a model to compute the optimal schedule of practice. *Journal of Experimental Psychology: Applied*, 14(2), 101–117.

Petrov, A., Dosher, B., & Lu, Z. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review*, 112(4), 715–743.

Philips, G. T., Kopec, A. M., & Carew, T. J. (2013). Pattern and predictability in memory formation: From molecular mechanisms to clinical relevance. *Neurobiology of Learning and Memory*, 105, 117–124.

Pimsleur, P. (1967). A memory schedule. *The Modern Language Journal*, 51(2), 73–75.

Pyc, M. A. & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. *Memory & Cognition*, 35(8), 1917–1927.

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.

Pyc, M. A. & Rawson, K. A. (2011). Costs and benefits of dropout schedules of test–restudy practice: Implications for student learning. *Applied Cognitive Psychology*, 25(1), 87–95.

Raaijmakers, J. G. (2003). Spacing and repetition effects in human memory: Application of the SAM model. *Cognitive Science*, 27, 431–452.

Raaijmakers, J. G. & Shiffrin, R. M. (1981). SAM: A theory of probabilistic search of associative memory. *The psychology of learning and motivation: Advances in research and theory*, 14, 207–262.

Raaijmakers, J. G. & Shiffrin, R. M. (2002). Models of memory. In D. Pashler, H. & Medin (Ed.), *Stevens' handbook of experimental psychology: Vol. 2. Memory and Cognitive Processes*. New York, NY: John Wiley & Sons.

Rawson, K. A. & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General*, 140(3), 283–302.

Reber, A. S. (1994). *Implicit Learning and Tacit Knowledge: An Essay on the Cognitive Unconscious*. Oxford University Press.

Rock, I. & Ceraso, J. (1964). Toward a cognitive theory of associative learning. In C. Scheerer (Ed.), *Cognition: Theory, research, promise* (pp. 110–146). New York, NY: Harper Row.

Roediger, H. L. & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17(3), 249–255.

Roediger, H. R. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, 59, 225–254.

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education*, 40, 4–17.

Rosch, E. & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4), 573–605.

Rumelhart, D. E. (1967). *The effects of interpresentation interval on performance in a continuous paired-associate task.* Technical Report 116, Institute for Mathematical Studies in Social Sciences, Stanford University.

Scardamalia, M. & Bereiter, C. (1996). Rethinking learning. In D. Olson & N. Torrance (Eds.), *The handbook of education and human development: new models of learning.* Wiley-Blackwell.

Schmidt, R. A. & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, 3(4), 207–217.

Shea, J. B. & Morgan, R. L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory*, 5(2), 179–187.

Shiffrin, R. M. & Schneider, W. (1977). Controlled and automatic human information processing: II. perceptual learning, automatic attending, and a general theory. *Psychological Review*, 84(2), 127–190.

Simon, H. A. (1966). A note on Jost's law and exponential forgetting. *Psychometrika*, 31(4), 505–506.

Simon, H. A. (1974). How big is a chunk? *Science*, 183(4124), 482–488.

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Education Review*, 29, 86–97.

Starch, D. (1927). *Educational Psychology.* New York: Macmillan.

Stigler, J. W., Fuson, K., Ham, M., & Kim, M. (1986). An analysis of addition and subtraction word problems in american and soviet elementary mathematics textbooks. *Cognition and Instruction*, 3(3), 153–171.

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. *Memory & Cognition*, 38(2), 244–253.

Thios, S. J. & D'Agostino, P. R. (1976). Effects of repetition as a function of study-phase retrieval. *Journal of Verbal Learning and Verbal Behavior*, 15(5), 529–536.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *Psychological Monographs: General and Applied*, 2(4).

Thorndike, E. L. (1913). *The psychology of learning*, volume 2. New York, NY: Teachers College, Columbia University.

Tsai, L. (1927). The relation of retention to the distribution of relearning. *Journal of Experimental Psychology*, 10(1), 30.

Underwood, B. J. (1964). Degree of learning and the measurement of forgetting. *Journal of Verbal Learning and Verbal Behavior*, 3(2), 112–129.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11), 1134–1142.

Valiant, L. G. (2006). A quantitative theory of neural computation. *Biological Cybernetics*, 95(3), 205–211.

Vaughn, K. E. & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science*, 22(9), 1127–1131.

Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: Does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review*, (pp. 1–7).

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, 39(5), 750–763.

Wickelgren, W. A. (1972). Trace resistance and the decay of long-term memory. *Journal of Mathematical Psychology*, 9(4), 418–455.

Wittgenstein, L. (1958). *Philosophical investigations*. Blackwell Oxford.

Woodson, M. I. (1974). Learner judgment in instructional decisions for learning meaningful paired associates. *Journal of Experimental Psychology*, 102(1), 167–169.

Wozniak, P. A. & Gorzelanczyk, E. J. (1994). Optimization of repetition spacing in the practice of learning. *Acta neurobiologiae experimentalis*, 54, 59–62.

Zeithamova, D. & Maddox, W. T. (2009). Learning mode and exemplar sequencing in unsupervised category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 35(3), 731–41.

Zhang, Y., Liu, R.-Y., Heberton, G. A., Smolen, P., Baxter, D. A., Cleary, L. J., & Byrne, J. H. (2011). Computational design of enhanced learning protocols. *Nature Neuroscience*, 15(2), 294–297.