Title
Making Experts: Optimizing Perceptual Learning in Complex, Real-World Learning Domains

Author
Thai, Khanh-Phuong

Publication Date
2015

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Making Experts: Optimizing Perceptual Learning in Complex, Real-World Learning Domains

A dissertation submitted in partial satisfaction

of the requirements for the degree of Doctor of Philosophy

in Psychology

by

Khanh-Phuong Thai

2015

ABSTRACT OF THE DISSERTATION

Making Experts:

Optimizing Perceptual Learning in Complex, Real-World Learning Domains

by

Khanh-Phuong Thai

Doctor of Philosophy in Psychology

University of California, Los Angeles, 2015

Professor Philip Kellman, Chair

How do we accelerate the process of gaining expertise? Recent research suggests that advanced pattern recognition and fluency can be developed in a short period of time using adaptive and perceptual learning technology (e.g., Kellman & Kaiser, 1994; Kellman, Massey, and Son, 2009). Much is still unknown, however, about the connections between perceptual learning and adaptive learning technology that allow for the efficient development of such expertise effects.

In six experiments, I examined a number of learning principles that bridge perceptual and adaptive learning and explored the generalizability of these learning principles across domains. In particular, I evaluated how different types of learning trial formats and feedback may bring about fluent structure recognition while improving training efficiency. To ensure that principles and experimental results are not confined to a single learning domain, I carried out these studies in two separate domains: mathematics and medical learning. Experiments 1, 3 and 5 trained undergraduates to interpret electrocardiogram recordings; Experiments 2, 4, and 6 replicated the

design of the other three experiments, but trained participants to map between graphical and symbolic representations of trigonometric and Exponential functions. Experiments 1 and 2 showed that the combination of passive exposure to the correct classifications and active classification practice enhanced fluency in pattern recognition while improving training efficiency. Experiments 3 and 4 explored the benefit of comparisons among contrastive examples and revealed that training with only comparisons can be detrimental, but that having some comparison practice can facilitate far transfer. Experiments 5 and 6 evaluated and demonstrated the effectiveness of a new paradigm that adaptively triggers paired-comparisons based on learners' error patterns to maximize training efficiency. Positive effects on learning were found in both learning domains.

These findings help to illuminate basic questions about the processes by which expert information extraction advances, and they inform our understanding of the general mechanisms that operate across learning domains. The results also lend themselves to applications in which learning interventions maximize the ease with which students pick up relevant structural relations in novel situations while minimizing training time.

The dissertation of Khanh-Phuong Thai is approved.

Robert A. Bjork

James W. Stigler

Sally Krasne

Noreen Webb

Philip Kellman, Committee Chair

University of California, Los Angeles

2015

TABLE OF CONTENTS

LIST OF FIGURES & TABLES

Experiment 1

Experiment 2

Experiment 3

Experiment 4

Experiment 5

Experiment 6

ACKNOWLEDGMENTS

Special thanks to Dr. Phil Kellman, who provided me with incredible mentorship. Thank you for your brilliance, generosity, and unwavering faith in me. Dr. Sally Krasne, thank you for your wisdom and support, and for those really fast and insightful email correspondences. Dr. Bob Bjork, Dr. Jim Stigler, and Dr. Reenie Webb, thank you for welcoming me to CogFog, TALL, and AQM. Your mentorship, insights, and kindness make me a better researcher and teacher. Thanks also to Dr. Ji Son for showing me how cool cognitive science is in the first place.

This dissertation was not possible without Joel Zucker and Tim Burke's impeccable code, Dara Afraz's graphic design flair, and Rachel Older's expert administrative support.

To my friends -- especially Everett, Genna, Veronica, Emma, and Carole -- thank you for the fun times, the good food, and the invigorating conversations about all things research and non-research related. Thanks also to everyone else in the Human Perception Lab, CogFog, TALL, and all the others who have contributed ideas and provided support throughout the years.

To my research assistants, Arthi, Blaise, Emma, Jake, Lawrence, Shriya, Siva, Xiaoya, and many more, thank you for all your help! I look forward to see what is in store for you!

To my parents, thank you for instilling in me the importance of education, and for your love and sacrifices. To my other Mom and Dad, Sen, Rose, Khuong, and Stephanie, thank you for your unconditional love and constant encouragement. Mason and Max, thank you for being so darn cute.

And to my husband, John---I could not have made it this far without your love, patience and support. You rock!

VITA

2007	B.A., Psychology Boston College Chestnut Hill, Massachusetts
2007-2009	Lab Manager, Human Perception Lab University of California, Los Angeles
2011	M.A., Psychology University of California, Los Angeles
2010-2014	Teaching Assistant and Teaching Associate Department of Psychology University of California, Los Angeles
2014	C. Phil, Psychology University of California, Los Angeles
2014-2015	Teaching Fellow Department of Psychology University of California, Los Angeles

PUBLICATIONS

Thai, K.P., Krasne, S., Kellman, P. J. (2015). The synergy of passive and active classification learning in electrocardiography. *Proceedings of the Annual Conference of the Cognitive Science Society* (July, 2015)

Yan, V. X., Thai, K. P., & Bjork, R. A. (2014). Habits and beliefs that guide self–regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory and Cognition.*

Thai, K. P., & Son, J. Y. (2013). The simple advantage in categorical generalization of Chinese characters. *Proceedings of the 35rd Annual Conference of the Cognitive Science Society* (July, 2013).

Thai, K. P., Mettler, E., & Kellman, P. J. (2011). Basic information processing effects of perceptual learning in complex, real–world domains. *Proceedings of the 33rd Annual Conference of the Cognitive Science Society* (July, 2011).

PRESENTATIONS

Thai, K. P. Krasne, S., Kellman, P. J. (July, 2015). The synergy of passive and active classification learning in electrocardiography. *Poster to be presented at the 37th Annual Meeting of the Cognitive Science Society* (Pasadena, CA)

Thai, K. P. Krasne, S., Kellman, P. J. (July, 2015). Perceptual learning with adaptively-triggered comparisons. *Poster to be presented at the 37th Annual Meeting of the Cognitive Science Society* (Pasadena, CA)

Thai, K. P., Zucker, J., Krasne, S., & Kellman, P. J. (October, 2014). The role of active classification training on learning to interpret electrocardiogram. *Poster presented at the Science of Learning in Medical Education Symposium* (Los Angeles, CA).

Thai, K. P., Massey, C., Kellman, P. J. (May, 2014). Perceptual Learning in Mathematics Education: Inverse Relations and Area Units. *Talk presented as part of a symposium on "Relational Understanding in Mathematics: Issues and Interventions" at the Association for Psychological Science meeting* (San Francisco, CA)

Thai, K. P., Son, J. Y., Kellman, P. J. (March, 2014). Perceptual learning in early mathematics: Emphasizing problem structure improves solving, mapping, and fluency. *Talk presented at the Society for Research on Educational Effectiveness* (Washington, D. C.)

Thai, K. P., Krasne, S., & Kellman, P. J. (October, 2013). Teaching pattern recognition in medical education: The combined effects of perceptual learning and declarative instruction. *Poster presented at the Science of Learning in Medical Education Symposium* (Los Angeles, CA).

Thai, K. P., Son, J. Y. (July, 2013). The simple advantage in categorical generalization of Chinese characters. *Poster presented at the 35th Annual Meeting of the Cognitive Science Society* (Berlin, Germany)

Yan, V., Thai, K. P., & Bjork, R. A. (April, 2012). Cultural differences in the self–regulation of learning. *Poster presented at the Western Psychological Association Convention* (Burlingame, CA)

Thai, K. P., Mettler, E., & Kellman, P. J. (2011). Basic information processing effects of perceptual learning in complex, real–world domains. *Talk presented at the 33rd Annual Conference of the Cognitive Science Society* (Boston, MA).

Thai, K. P. & Kellman, P. J. (May, 2011). Insights in Sight: Basic information processing effects of perceptual learning in complex, real–world domains. *Poster presented at the Vision Sciences Society meeting* (Naples, FL)

CHAPTER 1

Introduction and Overview

Experts differ from novices in fascinating ways. Experts are able to perceive patterns at a glance. For example, expert radiologists are able to detect cancer in a mammogram or a tumor in an *x*-ray in a split second (Kundel, Nodine, Conant, & Weinstein, 2007; Sowden, Davies, & Roling, 2000); an expert fisherman can easily distinguish between a pompano and a wahoo (Boster & Johnson, 1989); and chess grandmasters can spot an impending checkmate multiple moves in advance (de Groot, 1978), even without moving the eye (Reingold, Charness, Pomplun, & Stampe, 2001). Eye-tracking research has shown that experts focus faster and in greater proportion on relevant information while ignoring salient but irrelevant information (for a review, see Gegenfurtner, Lehtinen, & Saljo, 2011). In all of these domains, the important patterns – including relations that are quite abstract – are often invisible to novices, yet experts can recognize them rapidly and automatically.

The basis for such expert pattern recognition ability is *perceptual learning* (e.g., Gibson, 1969; Kellman, 2002). Defined by Eleanor Gibson (1969), perceptual learning refers broadly to experience-induced changes in the extraction of information (Gibson, 1969; Goldstone, 1998; Kellman & Garrigan, 2009; Kellman & Massey, 2013). Gibson also referred to perceptual learning as "differentiation learning", to emphasize the neural and perceptual changes in the way we encode information, resulting in our enhanced ability to detect and differentiate relevant information amidst variation. These processes are pervasive in perception and learning, allowing us not just to pick up of minute sensory details but also to extract abstract relations in complex, real-world learning domains (Kellman & Massey, 2013).

Despite its pervasive importance, perceptual learning (PL) has received little attention in instruction. A major part of the difficulty is that not just any kind of practice can foster PL. Indeed, PL often develops separately from formal instruction, and what is learned is often difficult to be verbalized. Furthermore, expertise is often thought to require some combination of maturation and many years of practice. A widely cited rule of thumb is that it takes at least ten years (Hayes, 1985), or 10,000 hours of diligent, "deliberate practice" (term coined by Ericsson, Krampe, & Tesch-Romer, 1993) with a task. This kind of description raises both theoretical and practical issues. In terms of theories of learning, it seems unlikely that the passage of time per se comprises the mechanism that generates expertise, nor is it plausible that all varieties of practice exert equal effects. More likely, learning advances due to particular variables in practice situations and their interactions with underlying learning mechanisms. In practical terms, this raises a critical question: how can we understand these variables and mechanisms in order to accelerate this process to support the development of real-world expertise?

Overview of Dissertation

The overarching goal of this dissertation was to examine important variables that may contribute to or enhance perceptual learning when combined with adaptive learning techniques. The studies aim to advance theoretical understanding, in terms of how learning works, and practical applications, in which instructional efforts attempt to develop expertise in real-world settings. In six experiments, we tested the benefits of active classification practice and passive exposure, of contrastive examples and of adaptively triggered comparisons for optimizing the fluent pattern recognition process in PL without sacrificing efficiency. At the core of our approach were major recent innovations in perceptual and adaptive learning technologies. We

begin with an overview of perceptual learning effects and a description of this innovative framework in Chapter 2. In Chapter 3, we provide an overview of the learning domains and the general methods used. In Chapter 4, we describe the effects of passive and active trial formats with Experiments 1 and 2. In Chapter 5, we explore the benefits of contrastive comparisons with Experiments 3 and 4. In Chapter 6, we study the additive benefits of adaptively triggered paired-comparisons with Experiments 5 and 6, and we conclude in Chapter 7.

In these experiments, we aimed to explore some general learning principles at the nexus of perceptual learning and adaptive learning to accelerate expertise. To ascertain that principles and experimental results are not confined to a single learning domain, we tested the same instructional manipulations in two unrelated domains: medical learning and mathematics learning. Experiments 1, 3, and 5 targeted the learning of electrocardiogram interpretation, and Experiments 2, 4, 6 examined the same effects in the learning of transformations in trigonometric and Exponential functions.

CHAPTER 2

Background

Perceptual Learning

Discovery and Fluency Effects

Kellman (2002) grouped perceptual learning effects into two major categories: *discovery effects* and *fluency effects*. With practice, we *discover* in a given domain what features and relations matter to important classifications (Gibson, 1969). If we keep at it, we also improve in *fluency*, or the ease and automaticity with which we extract these relevant features and relations (Kellman, 2002; Schneider & Shiffrin, 1977).

Fluency and discovery effects occur concomitantly, so improvement in one tends to lead to improvement in the other. For example, becoming selective in the use of information (a discovery effect) surely increases efficiency and improves speed (fluency effects). As we become more fluent at picking up relevant features or relations, we require less cognitive resources for that task, thus having more resources available to discover and process increasingly more complex features and feature relations, even those that are not initially evident (Bryan & Harter, 1899). Gibson (1969) emphasized these informational "invariants" - stable properties and relations that can lead to appropriate classifications. One example is recognizing a specific melody in different pieces of music that vary in scale and instrumentation. Such interplays of fluency and discovery effects pave the way for high-level thinking and complex problem solving (Kellman & Massey, 2013).

Indeed, issues of discovery and fluency in perceptual learning directly address barriers to transfer in all domains of expertise, including mathematics and science learning (Kellman & Massey, 2013). In mathematics, for example, students can memorize facts (i.e., a formula or a theorem) and the step-by-step procedures for solving a problem, but they often have trouble recognizing when those facts and procedures are appropriate, especially to a problem that doesn't look the same as those they trained with (e.g., Nunes, 1999; Givvin, Stigler, & Thompson, 2011; Stigler, Givvin, & Thompson, 2010). For example, a teacher may instantly recognize $x^2 - 3x + 2 = 0$ as a quadratic problem, equivalent to its transformation $(x-2)(x-1) = 0$. Students, however, often approach it as a "solve for x" problem — and get stuck trying to procedurally isolate x to one side of the equation.

As Kellman & Massey (2013) stated, "All effective use of declarative and procedural learning presupposes pattern recognition." To appropriately deploy relevant facts and procedures to new situations, students must efficiently recognize which facts and procedures are relevant to that situation (see also Joy Cummings & Elkin, 1999). That depends on how they classify the situation, which in turn depends on how they pick up information about the structure of the situation (e.g., the quadratic structure within the equation). The less effort they need to apply to these tasks, the better they are able to discover and process higher-order relations for critical thinking and problem solving in new situations (Kellman & Massey, 2013).

The crucial question in many educational domains is: how do we develop in students the fluent extraction of deep underlying principles from amidst irrelevant details, and the ability to transfer them from one situation to another situation that may look quite different from the first?

Perceptual and Adaptive Learning Modules

Until recently, the development of these perceptual learning has received little attention in instruction. Both familiarity with PL and suitable instructional methods have been lacking. Recent research, however, has shown that by leveraging natural perceptual processes to support advanced formal reasoning and transfer, PL can be systematically accelerated in real-world learning domains (e.g., Goldstone, Landy, & Son, 2008; Kellman & Massey, 2013; Kellman, Massey & Son, 2009). In our work, PL methods are realized in perceptual and adaptive learning modules (PALMs). PALMs incorporate a number of goals and features that have the potential to accelerate expertise. These go beyond the "10,000 hours" guideline for the development of expertise in a number of ways. Among these is that ordinary work flow may be haphazard and may not provide a full range of relevant problems, structures and information extraction demands; ordinary exposure or instruction may not provide enough exposure to challenging cases; in learning important classifications, it is important to incorporate two kinds of variability: variability in the incidental characteristics of exemplars that do fit in some category and variability in exemplars that do not fit in that category or comprise members of other categories; ideally, laws relating to spacing in learning should be incorporated, but these require special techniques (Mettler & Kellman, 2014); ordinary learning situations are seldom individually adapted to learners to offer beneficial spacing and to focus effort where it is most needed; and learning seldom tracks all classifications or categories to be learned to objective mastery criteria for each learner. The combination of perceptual and adaptive learning techniques in PALMs aims to address all of these issues to enhance and accelerate learning.

Perceptual Learning Features

In PALMs, students are often not asked to solve a problem, but rather to recognize what kind of problem it is. To do so, we engage students in making classifications or mappings of representations based on their underlying structures. We employ unique instances and systematic variations to tune learners' attention to relevant diagnostic structures. Learning to recognize the patterns unique to various structural categories (e.g., different diagnostic heart patterns observed on electrocardiograms) requires discriminating among a large number of exemplars from each category in order to extract the features common, albeit in variable form, to a given category and distinguishable from those of other categories. Furthermore, practice with mostly novel instances encourages spontaneous exploration and discovery of higher-order relations (e.g., Gibson and Pick, 2000, p. 169) across superficially dissimilar instances, instead of memorizing labels for particular instances.

To illustrate, in the Algebraic Transformation PALM, students view target equations and are asked to choose an equivalent equation, produced by a valid algebraic transformation, from among four choices. Students must answer quickly, so they experience many such trials while working through a module. The equations vary systematically and almost never repeat. This enables students to extract the underlying structural bases to make classifications of novel equations, and to transfer their knowledge to problem solving. What they learn is often generalizable to novel instances.

As a result, even though students never solve equations in this module, just a few hours of this kind of practice can lead to substantial improvements in fluency at solving algebraic equations. It was found that students took an average of about 28 seconds per problem before the

training and only 12 seconds per problem after the training (Kellman et al., 2008; Kellman, Massey & Son, 2009).

The success of this method at advancing students' fluency in grasping crucial structures and relations and detecting them in variable contexts has been shown in a number of other educational domains. We found that PALM interventions can accelerate fluent use of structure in contexts such as the mapping between graphs and equations (Kellman et al., 2008; Silva & Kellman, 1999), apprehending molecular structure in chemistry (Russell & Kellman, 1998; Wise, Kubose, Chang, Russell, & Kellman, 2000), understanding fractions and proportional reasoning (Kellman et al., 2009; Massey, Kellman, Roth, & Burke, 2011), discriminating between pathologic processes in skin histology images, identifying skin-lesion morphologies, and diagnosing wrist fractures in radiographs (e.g., Krasne, Hillman, Kellman, & Drake, 2013).

Adaptive Sequencing Features

This process of learning and discrimination is made more efficient by adaptive sequencing of trials that address different learning categories. The sequencing is based on an algorithm[1] that determines a "priority score" for each category, which in turn determines the sequence in which the next exemplar of each category is presented. The algorithm dynamically adjusts priorities after each learning trial based on learner accuracy and speed, as well as the number of trials since a category was last presented (Mettler, Massey & Kellman, 2011; Mettler & Kellman, 2014). As a result, learners spend most of their practice time with the classifications that they have most trouble with. This sequencing algorithm implements several principles that have been shown in learning research to strongly influence learning; these involve spaced

[1] 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.

practice, dynamics of short and long-term memory and relating recurrence intervals to underlying learning strength. One important principle is that as learning strength increases (as indicated by accurate and quicker responding), the delays in presenting problems that utilize the same concept or pattern (i.e. are in the same "category") should increase, with interleaved "interfering" problems (i.e., those from different "categories") occurring in between. This is in accordance with the "retrieval effort hypothesis", which states that more difficult but successful retrievals are more beneficial for learning (Karpicke & Roediger, 2007; Pyc & Rawson, 2009; Storm, Bjork, Storm, 2010).

The use of learning criteria that include a "target response time (RT)" within which a problem must be answered in order to be considered fluent is used for several reasons. First, fluent processing, indexed by accurate response under a target response speed, tends to indicate the operation of pattern recognition processes rather than lengthy conscious analyses. Secondly, attaining learning criteria including fluency may predict better retention of learning. Thirdly, fluency also implies reduced cognitive load, allowing the learner to perform important classifications in more complex and demanding contexts. Once a fixed number of successive exemplars of a category are answered accurately within the target RT, the category is "retired" so that future questions focus on exemplars from less well-recognized categories.

Recent results indicate that these adaptive algorithms outperform classic adaptive learning methods and non-adaptive presentation in both factual and perceptual learning (Mettler, Kellman & Massey, 2011; Mettler & Kellman, 2014). Such results are exciting, suggesting that PALMs can dramatically accelerate perceptual learning processes, and provide a needed complement to regular classroom instruction. Much is still unknown, however, about which training components prepare learners for the efficient development of such expertise effects in

real-world settings. Here we consider a few basic learning components.

CHAPTER 3

General Methods

In this chapter, we describe the learning domains and the experimental methods that were shared among the six experiments.

Learning Domains

Much of the research on perceptual category learning has been carried out under highly constrained and standardized conditions with artificial categories and stimuli. It is unclear how the nature of perceptual learning in these cases is generalized to realistic learning domains and across learning tasks. Even when experiments used realistic learning domains, we are often left not knowing how well the intervention applies to other learning domains.

Thus, to explore the domain-generality of the learning principles involved, we designed and tested two sets of PALMs to target learning in two different learning domains: one for electrocardiogram interpretation and one for recognition of Sine and Exponential transformations.

Medical: Electrocardiogram Interpretation

What are electrocardiograms (ECGs)?

ECG traces are recordings of tiny electrical changes on the skin that are caused when the heart muscle depolarizes during each heartbeat. A 12-lead ECG is one in which 12 different electrical signals are recorded at approximately the same time and are often used as a one-off recording of an ECG. It is one of the simplest and oldest cardiac investigations available, yet it can provide a wealth of useful information and remains an essential part of the assessment of cardiac patients.

Why ECG?

Visual interpretation of electrocardiograms (ECGs) requires superior perceptual recognition skills that often require years of practice to attain (Wood, Batt, Appelboam, Harris & Wilson, 2013; Salerno, Alguire, & Waxman, 2003; Mele, 2008), making it a great domain for perceptual and adaptive learning training.

It is a difficult skill to master for both medical students (Jablonover, Lundberg, Zhang, Stagnaro-Green, 2014) and doctors of different grades and specialties (e.g., Montgomery et al., 1994; Morrison & Swann, 1990; Gillespie, Brett, Morrison & Pringle, 1996; De Jager, Wallis, & Maritz, 2010). For example, the ability to correctly identify potentially life-threatening conditions was 57% in a group of US graduating medical students (Jablonover et al., 2014) and 46.4% in a group of South African Emergency Medicine trainees (De Jager et al., 2010). Computerized ECG interpretation software is now built into many modern ECG machines aimed to automate the process, but they show poor diagnostic accuracy with up to 46.5% error rates (Shah & Rubin, 2007; Bhalla, Mencl, Gist, Wilber & Zalewski, 2013; Ducas et al., 2012). It is clear that the current generation of computerized ECG interpretation technology should not be solely relied upon and its interpretation should be independently verified by an appropriately qualified individual (Estes III, 2013; cf. Fent, Gosai, & Purva, 2015).

One specific difficulty is with discriminating relevant from irrelevant information in ECGs. For any diagnostic pattern, some of the locations contain relevant information, while some do not. Each category involves patterns of diagnostic features, but the features are variable across the ECG traces. Salient features of an ECG trace do not necessarily indicate an abnormality, and waveforms that indicate normality on one lead may not be normal on another lead. Thus, learners have to know not only what to look for, but also where to look for them.

Despite such complexities, training with ECG PALMs with 3rd and 4th year medical students has shown promising results. Recent PALM studies suggest that medical students can learn to interpret 15 heart patterns with remarkable improvements from pretest to immediate posttest and maintained what they have learned for over a year later, all within just a few hours of practice (Krasne, Stevens, Kellman, & Niemann, under review)

ECG PALMs

We focused the ECG PALM training on 7 different heart patterns, one of which was normal (patterns of healthy patients without abnormalities). The materials for all PALMs consisted of 250 unique 12-lead ECG traces from real patients, with 26 - 46 unique traces for each of seven categorical diagnostic patterns. The diagnostic patterns we chose did not require participants to view the entire 12-lead ECG to make a classification, because each heart pattern contained relevant information in either the left half or the right half. For some heart patterns, half of the graph contained no useful information (e.g., the right half for LAD and RAD, left half for Anterior STEMI). Depending on the experiment, we used full 12-lead and half (right half and left half) images for each trace.

Mathematics: Transformation of Sine and Exponential Functions

Because of the strong algebra-geometry connection of this topic, we called these modules AlgGeo PALMs. The AlgGeo PALMs aimed to train participants to appropriately map between graphs and equations that represent the same mathematical function. Here we focused on the graphical and algebraic transformations of Sine and Exponential functions.

Earlier research (e.g. Silva & Kellman, 1999) showed that this makes a good testbed for manipulating variables relevant to perceptual learning. Even though students have been introduced to the topics covered in this training in high school, Silva & Kellman (1999) and our

recent data suggest that a majority have not mastered or have forgotten these skills[2]. For example, students may remember what the function $y = sin(x)$ looks like on a graph. When asked what the graph of $y = sin(x - 2)$ would look like (the same, but shifted to the right), they often have trouble connecting what they know to visualize the answer, even when they can recall the steps required to graph this function. Their difficulties are likely due in part to the limitations of traditional instruction and can be overcome by perceptual learning (e.g., Kellman, Massey, & Son, 2009; Landy & Goldstone, 2007; Kellman & Massey, 2013).

Mathematical representations are aimed at making concepts and relations accurate and efficient, but they pose complex decoding challenges for learners. Each representational type (e.g., a graph or an equation) has its own structural features and depicts information in particular ways. The rationale for the PALM was that fluent use of each representational type requires the ability to extract particular structural attributes (e.g., knowing where to look on a graph to obtain the information about the transformation, and to map that appropriately to an equation structure). Practice with mapping across representations requires accurate selection of information in each representational type and may also lead to intuitions about the way equivalent structures relate across representational types (e.g., learning the graphical consequences of shifting on the x-axis or scaling on the y-axis).

Furthermore, understanding of transformations provides a basis for later science learning and applications, as graphs and equations are pervasive in mathematics, physics, economics, and any other quantitative disciplines. This is also one of the key concerns in the common core state

[2] Perhaps with the exception of recent UCLA undergraduates, who in a pilot study were able to get about half of the questions correct at pretest. This prompted us to run the experiments on Amazon Mechanical Turk (MTurk). One prerequisite for participating in the study was that participants had passed Algebra II or an equivalent course (i.e., College Algebra), during which transformations of trigonometric and Exponential functions were generally introduced. Many MTurk participants had taken more advanced classes, but self-reported that it had been many years since they learned the materials.

standards (2010) for high school mathematics: students should be able to "interpret functions given graphically, numerically, symbolically, and verbally, translate between representations", and transfer their learning to "build new functions from existing ones".

AlgGeo PALM

The AlgGeo PALMs focused on four transformations of Sine and four transformations of natural Exponential functions from the canonical functions $y = sin(x)$ and $y = e^x$ or $exp(x)$: x-shifting, y-shifting, x-scaling, and y-scaling. Each of the transformation had 2 subtypes to account for the direction of the transformation. For example, the x-shifting category contains 2 subcategories: x-shifting to the left (e.g., $y = sin(x + 4)$) and x-shifting to the right (e.g., $y = sin(x - 4)$). Thus, each function family had 8 subcategories of transformation. There were 9 unique instances for each transformation subtypes[3], making a total of 144 images used in the training. The ranges on the axes varied across graph. This was necessary to create variation within each category, as well as to properly scale each function.

Similarities and Differences between These Two Domains

Both are excellent perceptual learning domains that require learners to extract relevant features and relations within highly abstract visual representations. Each category can be defined with a set of rules, but in both domains, there are many variations among instances of the same categories and many similarities between instances of different categories, so learners have to know what features or relations to look for, and where and how to look for them.

They differ in two important ways: ECG traces are relatively more complex and arguably contain more intricacies (extraneous information) than Sine and Exponential functions. Also, our

[3] This is with the exception of Experiment 2, in which the compression subtypes included only 4 instances, making a total of 119 unique graphs. This was because when the graphs within the compression subtypes were similarly scaled, they were very difficult to tell apart (e.g., $y = sin(x)/8$ and $y = sin(x)/9$). In later studies, we corrected for this issue by rescaling the graphs and added them back into the stimuli pool. Thus there were an equal number of instances per category subtype in Experiments 4 and 6.

participants have had more relevant background in mathematics than in ECG. These are noteworthy because prior research has shown that the degree of prior knowledge a learner has, and the degree to which stimuli include features that are irrelevant to the categorization task can influence the overall similarity of the exemplars, which contribute to learner's ability to extract relevant category information (e.g., Kalyuga, 2007; Gentner & Markman, 1994).

Our goal was to examine the effect of general learning principles on perceptual learning training. Thus, we expected that any learning effects from the training components (passive vs. active trial format, contrastive comparisons, and adaptive comparisons) would be similar in both domains. Any differences between the two domains, however, would raise interesting questions for future research on the importance of prior conceptual knowledge and the complexity of the stimuli on perceptual learning training.

Overview of Procedure

All experiments used a between-subject, pretest-training-posttest-delayed test design. The ECG experiments (1, 3, 5) included a primer prior to the pretest. Because the AlgGeo experiments (2, 4, 6) were conducted online, they included instruction check questions, practice trials, and extra survey questions following the immediate posttest and delayed test.

To investigate how different training components mediate perceptual learning, we manipulated the trial format and feedback within each PALM to create multiple versions for use in each experiment.

Dependent Variables of Interest

In a large respect, effective training for expertise means improving transfer and retention capability. We gave all participants a pretest before the training, and to gauge learning (rather

than performance, Soderstrom & Bjork, 2015) and retention, we gave them an immediate posttest right after the training and another posttest after a one-week delay. Thus we had three phases of assessment: pretest, immediate posttest, and delayed test. At these assessment phases, we were interested in the following three measures.

Efficiency

Since we were interested in optimizing our training, efficiency was our main measure. We considered two efficiency measures that took into account each learner's accuracy and the amount of training invested: (1) *trial efficiency*, as accuracy gain divided by the number of learning trials completed, and (2) *time efficiency*, as accuracy gain divided by the total time invested (accuracy per minutes of training).

Transfer Accuracy

Transfer, or the ability to use knowledge flexibly and effectively in new situations, is an important component of proficiency. To assess discovery gains from PALM training, we compared participants' accuracy in applying that they had learned to new situations. For ECG interpretation, transfer tests involved diagnosing novel ECG traces of trained patterns. In mathematics, near transfer tests involved mapping among graphs and equations of new Sine and Exponential functions. Far-transfer tests involved Cosine and Logarithmic functions and more complex combination functions, both of which measured participants' ability to extend learned transformations to new (but related) and more complex functions.

Fluency

In assessing enhanced pattern recognition (instead of analytic, reasoning processes), we also analyzed changes in fluent accuracy (defined as accurate responses made quickly, i.e.,

within the designated target RT). Thus, for fluent accuracy, we analyzed the pre-, post- and delayed test data, excluding responses that were not made within the 15 seconds target RT.

Survey Measures

Because our manipulations may produce not only differences in cognitive aspects of learning, but also in the motivational and engagement aspects, we asked participants to report their levels of engagement and enjoyment of the training experience, and to provide a judgment of learning and memory for the delayed test.

Overview of Analysis

Because we sought to compare differences across training conditions, we conducted planned comparisons among conditions. All statistical tests were two-tailed, with a 95% confidence level. Due to small sample size in each experiment, Bonferroni corrections for multiple pairwise comparisons can seriously raise Type II error (Perneger, 1998; Nakagawa, 2004). Thus, we followed the recommendations of Nakagawa (2004) and provided effect size estimates to evaluate the strength and direction of each relationship in our multiple tests. We reported effect sizes for ANOVA's using partial eta-squared (η^2_p) with .01 indicating a small effect, .06 indicating a moderate effect, and .14 indicating a large effect. In cases where there were differences at pretest, we conducted analysis of covariance (ANCOVA) with the pretest as the covariate to partial out the effect of the pretest. All assumptions for ANCOVA were met (i.e., pretest measures did not vary by condition, and there were never violations of the homogeneity of regression slopes, all p's > .10). Whenever there was good evidence that pretest variations were mostly due to chance, we also analyzed immediate posttest and delayed test data independent of the pretest.

To estimate the practical significance of differences between conditions, we computed

effect sizes (Cohen's *d*) as the difference in gain scores between conditions divided by the pooled standard deviation of the gain scores. Similarly, for within-subject differences in phase (when comparing performance among pre-, post-, and delayed test), we calculated effect sizes (also Cohen's *d*) as the mean difference divided by the standard deviation of the difference scores (Lakens, 2013). The thresholds for Cohen's *d* are .2 for a small effect, .5 for a moderate effect and .8 for a large effect, and 1.30 for very large effect sizes (Cohen, 1988; Rosenthal, 1996).

CHAPTER 4

The Synergy of Passive and Active Classification

INTRODUCTION

There is considerable evidence that the learning process and resulting representational structure of categories depend on the learning task (e.g., Love, 2002; Markman & Ross, 2003; Yamauchi & Markman, 1998). PALMs typically employ active classification practice, but does active classification better support perceptual learning (PL) than passive exposures to appropriate classifications? The little research done on this topic is not conclusive.

Active classification refers to learning tasks where the learners select a category label for a presented example and receive feedback that informs their perceptual, attentional and decision processes. Passive learning provides the same category membership information, but learners study the example and the category label without engaging in the choose-and-correct cycle. Active classification has also been known in the literature as supervised classification learning, discovery learning, or selection learning. The passive learning task has also been known as unsupervised observational learning, exposure learning, or reception learning.

Active versus Passive

The benefit of active retrieval over passive exposure has been well studied in memory literature. William James (1980) once wrote: "A curious peculiarity of our memory is that things are impressed better by *active* than by *passive* repetition" (p. 646, italics added). This reflects many findings in memory literature.

When learners actively engage with the learning material – by answering test questions about it, spacing out learning events, interleaving different learning items, or generating answers

– its representation in memory is changed such that the material becomes easier to recall in the future (e.g., Bjork, 1975). This improvement is often greater than that gained by repeated exposure to the same information (e.g., Bjork & Bjork, 1992; Roediger & Karpicke, 2006; Kornell & Bjork, 2008). The benefit of these active training conditions, called desirable difficulties (Bjork, 1994), was discovered in numerous experiments with humans who were trained in various verbal learning paradigms, such as the memorization of word pairs and prose materials.

Does active learning triumph over passive exposures in PL as well? Ordinary experience suggest that passive exposure alone can lead to discovery of relevant features and relations in PL. Children learn to tell dogs from cats by seeing a number of instances of dogs and cats. The novice participants in the classic chick sexing study by Beiderman and Shiffrar (1978) learned to categorize day-old baby chickens with just a single page of instruction. Novice wine drinkers can learn to discriminate between wines without any instruction (Hughson & Boakes, 2009). People can learn to recognize the styles of artists in new paintings by passive viewing of multiple samples of each artist (Kornell & Bjork, 2008).

In some cases, passive presentations may actually be better than active presentations. Passive presentation in the form of worked examples is the preferred mode of learning for novices (e.g., Recker & Pirolli, 1995), and is an effective instructional alternative to solving problems in a variety of domains. Paas and van Merrienboer (1994) studied student learning of geometrical problem solving skills and found that when students studied worked examples of problems (*passive*), they attained better accuracy on solving new problems than those who had to solve problems from scratch (*active*). Paas and van Merrienboer postulated that a considerable part of the mental effort in the active condition was allocated to processes that were irrelevant for

learning. Those in the *passive* condition, on the other hand, could focus on the relevant aspects of problem structure and solutions, thus requiring less training time and less mental effort. *Passive* learning trials also offer error-free exposures to the classifications to be learned, eliminating residual effects of incorrect guesses that may occur in *active* learning. When studying worked examples, the learner is freed from performance demands and s/he can concentrate on gaining understanding (e.g., Renkl, Atkinson & Grobe, 2004). This passive exposure to the classifications to be learned can facilitate subsequent perceptual encoding processes involving those classifications (see also Jacoby, Toth, Lindsay, & Debner, 1992). Conversely, Bodemer & Faust (2006) found that when asking students to make active connections between multiple representations of fractions, they were better able to understand the underlying structures of fractions than when they passively observed the correspondences.

Much of the category learning literature focuses on classification, how one learns to assign instances to categories. However, category learning is not simply classification learning. How the categories were learned is likely to have a large influence on how the category is represented (e.g., Anderson, Ross, & Chin-Parker, 2002). Indeed, an active task tends to encourage learners to focus on information that distinguishes categories, while a passive task tends to engage them with finding within-category regularities (e.g., Markman & Ross, 2003; Chin-Parker & Ross, 2002; Carvalho & Goldstone, 2014). In a recent article, Levering and Kurtz (2015) compared the category knowledge produced by an active classification task and a passive observational learning task. They trained participants to discriminate between two artificially created categories, each with 5 stimuli, in which a single feature determined category membership and other features correlated but did not perfectly predict category membership. They found that the active learning task biased learners toward more discriminative learning

compared to the passive learning task. However, passive learning allowed for enhanced sensitivity to the features that were not perfectly predictive. Their reasoning was, active classification requires explicit comparison and weighing of multiple category options of each trial, which emphasizes the mutual exclusivity of categories and encourages hypothesis testing about the diagnosticity of particular features. Just having passive exposure to the correct classifications, on the other hand, may support a broad understanding of the coherence among members of a category (see also Hoffman & Murphy, 2006, and Hsu & Griffiths, 2010).

Combining Passive Exposure with Active Classification

Since passive and active processes have complementary benefits, it is possible that combining passive and active learning may be most beneficial. One can imagine everyday life situation in which a domain expert must successfully perform many repetitions of a particular discrimination, but that s/he must first have as the basis for comprehension and inference robust concepts that capture the nature of each category, not just how to tell any two categories apart. This hypothesis accords with research on skill acquisition by Renkl, Atkinson, and colleagues under the ACT-R framework (e.g., Atkinson, Derry, Renkl, & Worthham, 2000; Kalyuga, Ayres, Chandler and Sweller 2003; Renkl, Atkinson & Grobe, 2004), in which passive study of examples is valuable early in training. Much of this work focused on procedural problem solving domains, for which a smooth transition (fading) from study of worked-out examples to problem solving may be ideal. Initial passive presentations can reduce cognitive load early in training when it is highest by not having to engage in decision-making processes, resulting in fewer unproductive learning events (Renkl and Atkinson 2003; Renkl, Atkinson, Maier, & Staley, 2002). Active learning, in contrast, forces guessing at the start, which might lead to cognitive overload. Wrong guesses or hypotheses may also tend to linger and impede later learning. In

addition, being forced to produce responses without knowing much may be frustrating, undercutting motivation in some learners.

Most potential advantages of passive exposure can be realized by using passive trials only at the start of learning. Initial passive study in PL might focus learners' attention on specific features that define each category and in turn support the acquisition of the category representation. As the learning progresses, active learning can support discriminative processes needed for correct classification. Active learning after an initial stage may be especially valuable in an adaptive framework. We sought to test this hypothesis in two real-world, complex PL domains.

Overview of Experiments 1 and 2

In Experiments 1 and 2, we asked: (1) Does the active classification experience enhance structure recognition? (2) Does the combination of passive exposure and active classification improve training efficiency?

In Experiment 1, we trained undergraduates to classify seven diagnostic patterns in electrocardiography. In Experiment 2, we trained Amazon Mechanical Turk workers to identify the transformations in graphs and equations of Sine and Exponential functions. In each experiment, we created three versions of the PALM involving: (1) only *active* classification to the underlying diagnostic pattern, (2) only *passive* presentations of the correct classifications, (3) initial passive presentations followed by active classifications (*passive-active* condition). The *active* and *passive-active* conditions involved classification with feedback and were adaptive to the learner's performance, and the *passive* training involved study of the correct interpretations and was not adaptive. To compare learning across conditions, we examined participants' ability to correctly and quickly classify novel instances into trained categories of diagnostic patterns. All

active trials used an adaptive learning system – the ARTS (Adaptive Response-Time-based Sequencing) system (Mettler & Kellman, 2014).

Experiment 1

METHOD

Participants

90 undergraduate students (mean age = 19.75, 68 Female) from University of California, Los Angeles participated in this experiment for research credits. Of the 90 participants, we removed 3 participants from the *active* condition and their yokes from the *passive* condition because: one participant in the *active* condition self-reported to be "not at all engaged" with the module, one participant in the *passive* condition dropped out of the study after the Primer, and one pair because both participants did not return for the delayed test. We also removed 3 participants in the *passive-active* condition for different reasons: one reported to not having read the primer, one took notes during the primer and studied it before the delayed test, and one was not fluent in English and claimed to have had trouble understanding the primer. Of the 81 participants, 69 (23 in each condition) completed their assigned modules (either by reaching learning criteria or were yoked to those who did) within the time allotted.

Design

There were three between-subject training conditions to which participants were randomly assigned: (1) *active* PALM in which all learning trials were interactive, adaptive and with feedback, (2) *passive* PALM in which all learning trials were static and contained the correct descriptions and labels, and (3) *passive-active* PALM which contained a subset of passive trials prior to the active PALM.

Materials

The training consisted of two phases: a brief *primer* to ECG interpretation (same for all conditions) and the PALM phase with either *active*, *passive*, or *passive-active* task formats.

Primer. The primer consisted of a series of PowerPoint slides consisting of a brief explanation of the 12 ECG leads, how to measure widths and heights on the grid, and one example of a typical ECG trace for each heart pattern. In each example, the relevant features were marked and described, similar to samples provided in textbooks. No other information about the heart anatomy, physiology, or other basics of ECG interpretation was provided in the primer. *Appendix A.1* contains sample primer slides.

Quiz. The primer quiz asked participants to match the descriptions of the diagnostic features to each of the seven heart patterns shown on the primer. This was to ensure that participants were familiar with the diagnostic features of each heart pattern. *Appendix A.2* contains this quiz.

PALMs. The materials for PALMs consisted of 250 unique 12-lead ECG traces from real patients, with 26 - 46 unique traces for each of seven categorical diagnostic patterns. The seven patterns were: Normal, Acute Anterior ST Segment Elevation Myocardial Infarction, Acute Inferior ST Segment Elevation Myocardial Infarction, Right Bundle Branch Block, Left Axis Deviation, Right Axis Deviation, Old Inferior Myocardial Infarction. The following list contains the diagnostic features of each heart pattern:

1) *Normal*: no abnormalities
2) *Anterior STEMI* (Acute Anterior ST Segment Elevation Myocardial Infarction): ST elevation > 2 mm in two consecutive V1-V3 leads

3) *Inferior STEMI* (Acute Inferior ST Segment Elevation Myocardial Infarction): ST elevation >1 mm in II, III and aVF with ST depression > 1 mm in leads I and aVL
4) *RBBB* (Right Bundle Branch Block): QRS ≥ 0.12s and rsR' or rSR' (i.e. "rabbit ears") in V1 & V2; deep reciprocal S waves in left lateral leads
5) *LAD* (Left Axis Deviation): R net positive in I and net negative in II and aVF
6) *RAD* (Right Axis Deviation): QRS in I is negative and aVF is positive, or QRS is evenly divided in I with III more positive than aVF
7) *Old Inferior MI* (Old Inferior Myocardial Infarction): Significant Q's (0.04 sec & > ¼ the height of R) in at least two of II, III, and aVF; no ST elevation

In the *active* PALM, on each trial, participants chose among seven choices the diagnostic category for a given ECG trace. *Figure 1.1a* shows an example trial. Accuracy and speed were continually tracked; trial feedback was given after each response and block feedback was given after every 12 trials. The trial feedback played a sound corresponding to the correctness of the response, and displayed the correct answer, and response time when correct. It also marked relevant features on the ECG, along with a brief description of those features as seen in the primer. Block feedback provided mean accuracy and speed by block and percentage of categories completed. Feedback screens were not timed. *Figure 1.1b* shows an example feedback screen following an incorrect response. Categories were adaptively sequenced based on both accuracy and response times as according to the ARTS sequencing algorithm (see Mettler & Kellman, 2014). Categories were dropped (retired) from the training set after reaching learning criteria (i.e., correctly identified consecutively in 4 out of 4 presentations, each in under 15 seconds). Participants completed the module when they had retired all 7 categories.

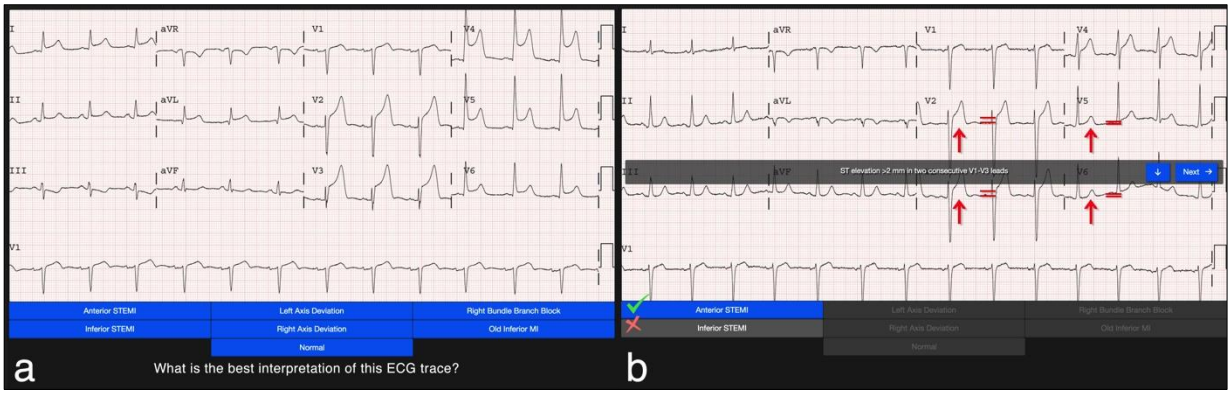


Figure 1.1. (a) Sample *active* classification trial; (b) Feedback provided when incorrect

In the *passive* PALM, each trial was the same as the correct trial feedback screen for the *active* group (*Figure 1.2*). The correct label, the relevant features and their descriptions were provided, and participants were asked to pay attention and to study each correct diagnosis. The *passive* condition thus did not have classification feedback and was not adaptive. To equate the total number of trials across two groups, we yoked each participant in the *passive* training condition to the total number of trials seen by another participant in the *active* training condition. To determine how many items per category to show, we used the average proportions of trials per category that a pilot group of *active* participants needed to complete the module. These proportions were similar across *active* participants, so we used the same proportions for all *passive* participants. The duration of each *passive* trial was 13 seconds, determined from the average amount of time it took pilot participants in the *active* group to respond and view the trial feedback. After 13 seconds, the screen cleared. To keep the participants engaged and to equate the existence of a motor response with the *active* condition, participants had to click on a Next button to see the next trial, and there was a sound played to signal the beginning of each trial. There was an untimed break every 12 trials.

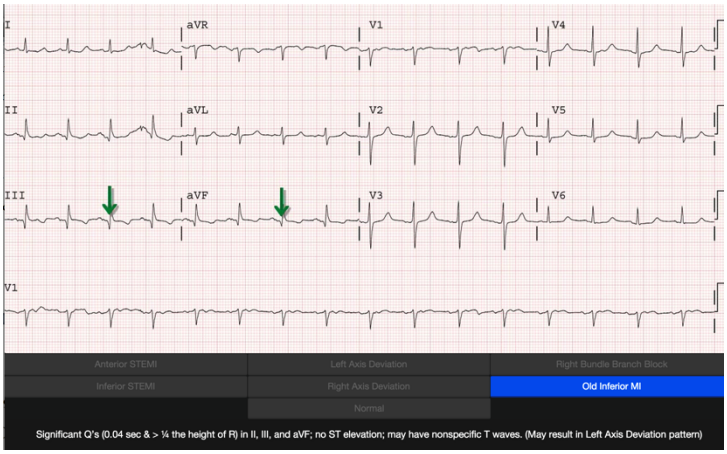


Figure 1.2. Sample *passive* trial.

In the *passive-active* PALM, participants viewed a set of 14 *passive* trials (two examples from each category) as in the *passive* condition, in random order, before moving on to the adaptive *active* classification trials for which participants received the same feedback and learning criteria as those in the *active* classification condition. All three PALMs used the same pool of ECGs.

The *active* and *passive-active* groups received adaptive training and learned toward learning criterion, while the *passive* group received one-to-one yokes of the number of trials and durations per trials for the *active* group to reach learning criterion. Thus the *passive* group did not receive adaptive training.

Assessments. Three assessments, each consisting of 14 new ECG's (two from each category), were used in counterbalanced order as pretest, posttest, and delayed test. None of the ECGs used in the assessments appeared in the PALM. Each assessment trial presented an ECG and seven answer choices (*Figure 1.1a*). No feedback was given after each trial.

Survey. The survey asked about their prior knowledge of ECG reading, levels of engagement and enjoyment of the training experience, judgment of learning and memory,

amount of sleep they had the night before, four questions to assess their intrinsic theory of intelligence (from Chiu, Hong, & Dweck, 1997), demographics information (age, gender, college year and major, English fluency), and general comments about their experience in the study. *Appendix B.1* contains the full survey. The sequence of questions was the same for all participants.

Procedure

Figure 1.3 displays the procedure of this study. Participants were given 20 minutes to study the *primer* followed by a quiz on which they were asked to match the descriptions of the diagnostic features to each of the seven heart patterns shown on the primer. They checked their answers afterward.

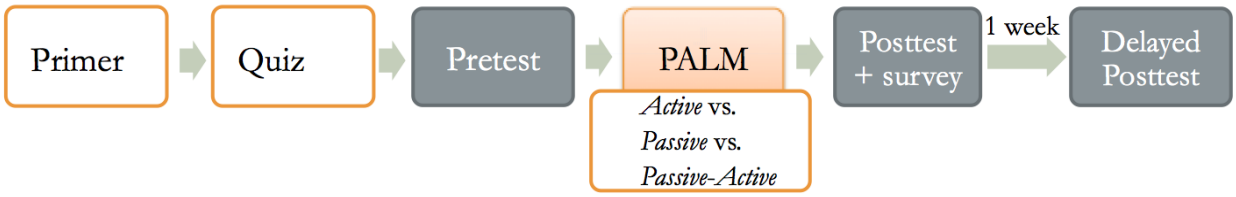


Figure 1.3. Experiment 1 procedure.

After the quiz, participants took the pretest and were randomly assigned to learn with the *active*, *passive*, or *passive-active* PALM. When participants finished the module (or after the 2-hour time allotted), they completed the immediate posttest and a survey. Participants returned for the delayed-posttest one week later.

Overview of Analyses and Expected Results

Adaptive learning technology aims to bring learners to a learning goal, so to assess condition differences, we compare performance of learners who completed their assigned

PALMs. Thus, we analyzed performance from those in the *active* (and their *passive* yokes) and the *passive-active* groups who reached learning criteria during the time allotted. It is important, however, to consider performance from all participants who have attempted the module and compare their completion rate, to understand how training conditions differentially bring learners to mastery criteria. We report data from participants who completed the PALM in the main text and report the data from all participants in *Appendix C.1*. Generally, the same patterns of results were found when we included all participants who did not reach learning criteria. It is likely that given more time, the remaining participants would also have reached learning criteria.

Participants in the *active* group on average retired 87.3% and the *passive-active* 89.9% of the categories. Four participants from each condition did not complete the assigned modules (out of 81, 23 per condition did).

Based on prior work, we expected all PALMs to produce robust improvements in classification. We hypothesized that the *passive-active* group would produce the best results. Because we used learning to criterion, our primary measure was learning *efficiency*, defined as accuracy gain from pretest to posttest divided by the number of training trials invested. We expected the *active* group to have greater improvements in accuracy and/or response time (for correct answers - RTc) than the *passive* group. At pretest, participants showed slight differences in pretest accuracy between conditions. Though these differences were not statistically significant at $\alpha = 0.05$, pretest accuracy could influence the amount of posttest gained from the training. Thus, aside from using analysis of variance (ANOVA) to confirm condition differences on raw scores, we also ran analysis of covariance (ANCOVA) on accuracy gain, fluent accuracy gain, and efficiency with pretest performance as a covariate. This allowed us to control for the effect pretest had on the posttest gains. All assumptions for ANCOVA were met for each

dependent variable: (1) independence of the covariate and the treatment effect (the pretest were not different across groups), $F(2,88) < 2$, p's > .10, (2) and (2) homogeneity of regression slopes (the correlation between pretest and the posttest gains were roughly equal across conditions), $F(2,75) < 1$, p's > .10.

Yoking by number of trials was not perfect for 5 pairs of participants; however, we retained them in the analyses because (1) removing them did not change the results, and (2) total trials and training times were similar between the *active* and *passive* groups. The three groups did not differ on quiz performance or any other measures not reported here.

RESULTS

Efficiency

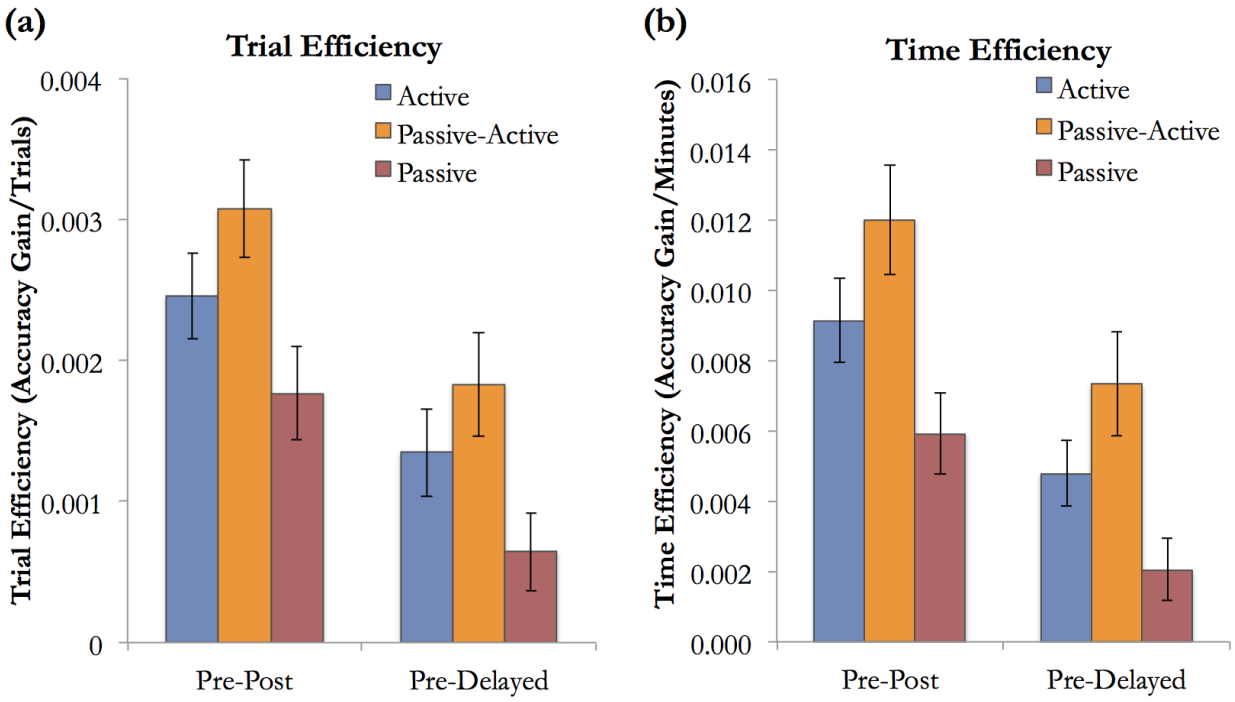


Figure 1.4. Efficiency (a) by trial and (b) by time. Error bars are ± 1 standard error.

Efficiency by Trials

Figure 1.4a gives efficiency as computed by trials. A 2 phase (pre-post, pre-delayed) x 3 condition (*active, passive, passive-active*) ANCOVA with pretest accuracy as a covariate confirmed that after controlling for the effect of pretest accuracy, there was a reliable main effect of condition, $F(2, 65) = 7.19$, $p < .01$, $\eta^2_p = .18$. There were no reliable differences between *active* and *passive-active* groups in the mean efficiency ($t(44) = 1.13$, $p = .26$), but both the *active* and *passive-active* groups had better efficiency than the *passive* group with medium to large effect sizes (.002 and .003 vs. .001, respectively, $t(44) = 2.05$, $p = .045$, $d = .62$, and $t(44) = 3.00$, $p = .004$, $d = 1.01$). The drop in efficiency from immediate posttest to delayed test was marginally reliable, $F(1, 65) = 3.31$, $p = .07$, $\eta^2_p = .05$. There was no phase x condition interaction, $F(2,65) = .04$, $p = .96$, $\eta^2_p = .001$.

Pretest accuracy was significantly related to efficiency scores, $F(1, 65) = 29.28$, $p < .001$, $\eta^2 = .31$, with pretest accuracy negatively correlated with both pre-post efficiency, $r(69) = -.42$, $p < .001$ and pre-delayed post efficiency, $r(69) = -.50$, $p < .001$. This suggested that pretest variations were largely due to chance. Thus, we also analyzed efficiency uncorrected for pretest variations (post- or delayed test accuracy/number of trials). Across both post and delayed tests, the *passive-active* condition outperformed the *active* condition with a medium effect size (.005 versus .004, respectively), $t(44) = 2.03$, $p = .048$, $d = .58$. *Passive-active* robustly outperformed *passive* with a large effect size, $t(44) = 3.83$, $p < .001$, $d = 1.12$. *Active* and *passive* did not differ reliably in overall efficiency uncorrected for pretest scores, $t(44) = 1.56$, $p = .13$.

Efficiency by Time

Figure 1.4b shows the efficiency as computed by time. Time efficiency showed the same patterns of results, with one exception. Unlike trial efficiency where there was no difference between the *active* and *passive* conditions on efficiency uncorrected for pretest variations, the

active condition produced higher time efficiency than *passive* with a large effect size, $t(44) = 2.73$, $p = .009$, $d = .82$. More details of these analyses are in *Appendix C.2*.

Accuracy

Figure 1.5a shows the mean accuracy by conditions. As expected, participants from all conditions produced strong learning gains. A 3 phase (pre, post, delayed test) x 3 condition ANOVA on accuracy confirmed a main effect of phase, $F(2,132) = 104.49$, $p < .001$, $\eta^2_p = .62$. Across all conditions, participants produced strong learning gains from pretest to immediate posttest (29% to 64%, $t(68) = 13.89$, $p < .001$, $d = 2.20$) and to delayed test (48%, $t(68) = 7.08$, $p < .001$, $d = 1.17$) with large and very large effect sizes. There was also reliable forgetting between immediate posttest and delayed test, $t(68) = 7.38$, $p < .001$, $d = .97$.

There were also differences in overall accuracy as a function of condition, $F(2,66) = 4.61$, $p < .05$, $\eta^2_p = .12$. The *passive-active* condition outperformed both the *active* and *passive* conditions on overall accuracy with medium effect sizes, $t(44) = 2.64$, $p = .02$, $d = .78$, and $t(44) = 2.15$, $p = .04$, $d = .62$, respectively. *Active* and *passive* did not differ reliably, $t(44) = 1.23$, $p = .22$. There was a marginally significant phase x condition interaction, $F(4,132) = 1.99$, $p < .10$, $\eta^2_p = .06$. There were no condition differences at pretest (p's > .10), but the *passive-active* group outperformed the *passive* group at both posttests (70% vs. 57%), $t(44) = 2.37$, $p < .05$, $d = .70$, and delayed test (50% vs. 39%), $t(44) = 2.75$, $p < .01$, $d = .82$, with medium-large effect sizes. The *active* group did not differ reliably from the *passive* group at immediate posttest, $t(44) = 1.48$, $p = .15$, but had marginally higher delayed test accuracy than the *passive* group with a medium effect size, $t(44) = 1.73$, $p = .09$, $d = .51$. The *passive-active* group had numerically higher means than *active* at immediate posttest and delayed test, but these differences were not

statistically reliable at immediate posttest, $t(44) = 1.22$, $p = .23$ or at delayed test, $t(44) = 1.59$, $p = .12$.

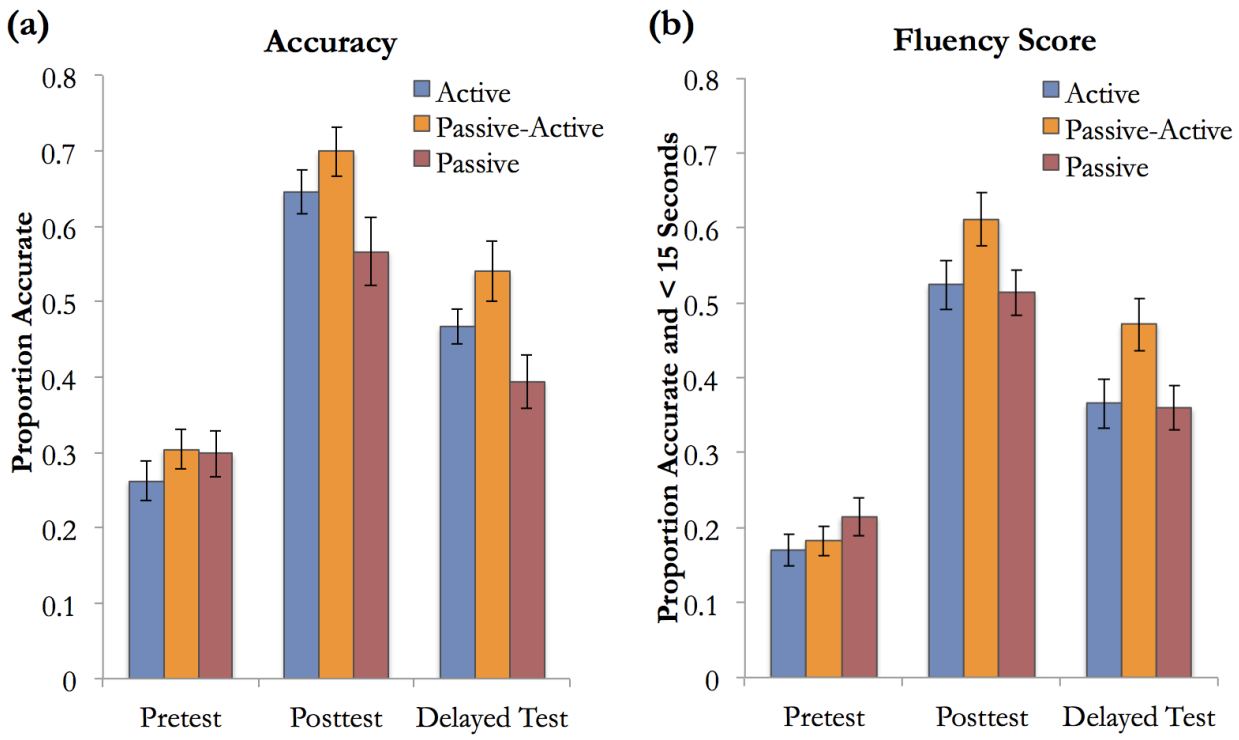


Figure 1.5. Mean (a) accuracy and (b) fluent accuracy. Error bars are ± 1 standard error.

Accuracy Gain

The pattern of condition differences was slightly different with accuracy gains. The *active* and *passive-active* conditions produced higher gains than the *passive* condition with medium effect sizes, $t(44) = 2.18$, $p = .04$, $d = .64$, and $t(44) = 2.41$, $p = .02$, $d = .71$, respectively. There were no reliable differences in accuracy gains between the *passive-active* and *active* conditions, $t(44) = .21$, $p = .84$, and no significant interactions, p's > .10.

Fluency

Figure 1.5b displays the fluent accuracy by condition. Fluent accuracy showed the same pattern as accuracy. The *passive-active* group tended to show higher overall fluent accuracy than

both the *active* (43% vs. 38%), $t(44) = 2.33$, $p = .03$, $d = .69$, and *passive* groups (43% vs. 38%), $t(44) = 2.23$, $p = .03$, $d = .66$, with medium effect sizes. Interestingly, *active* also outperformed *passive* on overall fluent accuracy gain. This was marginally significant but had a medium effect size, $t(44) = 1.72$, $p = .09$, $d = .51$. There were no other reliable condition differences. *Appendix C.2* contains more details of these analyses.

Progression of Learning

Because we wanted to examine the benefit of *passive* training trials prior to the *active* classification trials in the training, we compared accuracy between the *passive-active* and *active* conditions during the course of training. Participants varied in the total number of trials needed for reaching learning criteria, so we also compared their accuracies by quartiles of the training. *Figure 1.6a* shows the mean accuracy during the training over the 4 quartiles and *Figure 1.6b* over the first 12 training blocks. The *passive-active* group was reliably more accurate during the training than the *active* group, with a large effect size, $t(44) = 3.84$, $p < .001$, $d = 1.13$. This held true across all four quartiles with medium-large effect sizes, $t(44) > 2.21$, $p = .02$ to $.03$, $d = .77$ to $.89$. Interestingly, this superiority of the *passive-active* group over *active* group did not appear until the 4th training block, $t(52) = 2.95$, $p = .005$, $d = .80$. These two groups had similar RTc and fluent accuracy, p's $> .10$.

As a result, the *passive-active* condition required 30 fewer trials to reach learning criteria than the *active* condition, $t(44) = 2.19$, $p = .03$, $d = .65$. They also spent about 6 minutes less than the *active* condition (38 minutes vs. 44 minutes) but this difference was not statistically reliable, $t(44) = 1.45$, $p = .15$, $d = .43$. There were no differences in trials or time between the *active* and *passive* conditions, p's $> .20$. *Table 1.1* contains the descriptive statistics from the training for each condition.

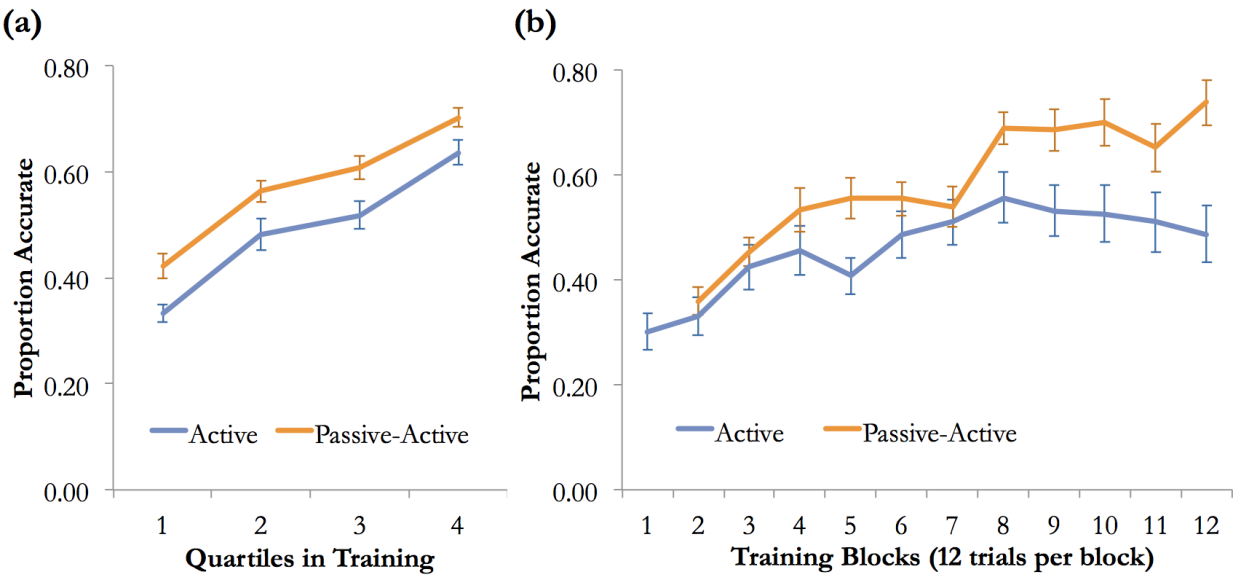


Figure 1.6. Mean accuracy in the training (a) by quartiles and (b) by blocks. Note that many participants did not experience all 12 blocks of training while some did and more to reach learning criteria. Each block contained 12 trials. The *passive-active* group received 14 *passive* trials in block 1. Error bars are ± 1 standard error.

Condition	Trials Completed	Minutes on Training	Training Accuracy	Training RTc	Training Fluent Accuracy
Active	167.5 (11.8)	43.96 (3.37)	.49 (.02)	7.58 (.38)	.46 (.02)
Passive-Active	137.8 (6.7)	37.73 (2.68)	.57 (.01)	8.10 (.48)	.52 (.01)
Passive	159.7 (8.6)	47.91 (2.40)	-	-	

Table 1.1. Training means by condition. Standard errors are in parentheses.

Self-Report Ratings and Judgment of Memory

On the survey, we asked participants to self-report "How enjoyable the training as a whole, on a scale from 1-6 (1 = not at all enjoyable, 6 = very enjoyable)". Our groups differed marginally in how enjoyable they rated the training, $F(2,61) = 2.51$, $p < .10$. The *passive-active*

PALM was found to be more enjoyable (M = 4.55, SD = 1.23) than the *passive* PALM (M = 3.76, SD = 1.22), $t(47) = 2.01$, $p = .05$, $d = .64$, and marginally more enjoyable than the *active* PALM (M = 3.90, SD = 1.14), $t(46) = 1.74$, $p = .09$, $d = .55$, both with medium effect sizes. There was no difference between the *active* and *passive* groups, $t(40) = .39$, $p > .70$. Those in the *passive-active* group also self-reported to be more highly motivated and engaged during the module (on a scale from 1-6, 1 = not at all, 6 = very much, M = 4.90, SD = .72) than the *active* (M = 4.38, SD = 1.02) and the *passive* groups (M = 3.95, SD = 1.36), $t(39) = 1.87$, $p = .07$, d = .59, and $t(39) = 2.77$, $p = .008$, $d = .87$, with medium to large effect sizes. There was no difference between the *active* and *passive* groups, $t(40) = 1.15$, $p = .26$.

The *passive-active* group also gave marginally higher ratings to "On a scale from 1-6 (1 = not at all, 6 = very much), how helpful was the training module?" than the *passive* group with a medium effect size (M = 4.67, SD = 1.16, vs. *passive-active,* M = 5.25, SD = .72), $t(39) = 1.93$, p = .06, $d = .60$, but not higher than the *active* group (M = 4.90, SD = 5.25), $p > .10$. There was no difference between the *active* and *passive* groups, $t(40) < 1$, $p = .45$. There were no other group differences on the remaining survey items, p's > .10.

DISCUSSION

The *passive-active* condition in this study, consisting of initial passive exposure, followed by active adaptive learning, produced better learning and transfer than *active* and *passive* learning for the same amount of time. It was also more enjoyable than the *passive* condition and marginally so than the *active* condition. *Passive-active* outperformed *active* adaptive learning on comparisons during the course of learning (*Figure 1.5*), as well as in accuracies and efficiencies uncorrected for what were likely random pretest variations across groups. Effect sizes for

learning differences between *passive-active* and *active* ranged from around .6 to .8, which are medium to large effect sizes.

The initial passive exposure speeds learning relative to starting with active classification, despite there was similar number of learning trials in the passive portion and the first active trial block. In the first few training blocks, the abrupt change from passive to active introduced similar error rates as those in the *active* group. However, from about the 4th block on, those in the *passive-active* group made fewer errors. These gains appear to be preserved through the course of learning and in posttests.

Active did not differ from *passive* on raw accuracy and fluency, but *active* produced higher accuracy gain, fluent accuracy gain, and efficiency (corrected for pretest variations) than the *passive* condition. The *active* condition in this experiment, as well as the active part of the *passive-active* condition, utilized the ARTS adaptive learning algorithm previously found to be highly effective in earlier work. The *passive-active* condition here appears to enhance a learning approach that has been previously shown to outperform classic adaptive learning systems and a number of presentation schemes in adaptive PL (Mettler & Kellman, 2014).

Experiment 2

The purpose of Experiment 2 was to examine the generalizability of the advantage of passive-active training task with a different learning domain involving transformations of Sine and Exponential functions and with a different, more diverse sample of participants. This use of a new learning domain and a different subject sample represents a robust test of the learning principles. If *passive-active* is an effective training in general, we should be able to replicate the

effect in Experiment 2. If the effects of Experiment 1 were specific to the learning domain, then Experiment 2 may not show the same effects.

METHOD

Participants

75 Amazon Mechanical Turk workers[4] (36 Female, mean age = 34.34, *SD* = 10.24) from the United States who have passed Algebra 2 or an equivalent course (i.e., College Algebra) completed the study online from their own computers. Each received $13 in remuneration. *Appendix D.1* contains a summary of other demographic information.

Design

There were three between-subject training conditions: (1) *active* classifications of graphed functions and their symbolic expressions, (2) *passive* exposures to the correct mappings only, (3) a combination of a set of passive trials containing the correct mappings, followed by active classifications (*passive-active*).

Materials

Instruction Check. Because participants self-administered the study, care was taken to ensure that they have read and understood the instructions before moving on to each phase. Following each set of instructions, there were 3-5 multiple-choice instruction check questions. Participants had to answer all instruction questions correctly before moving on, and were able to review the instructions after each incorrect try.

[4] There were initially 98 MTurk workers who participated in the study, 16 of whom (9 from *active*, 3 from *passive*, and 4 from *passive-active*) withdrew from the study during the training, and 5 did not return for the delayed test (1 from *active*, 1 from *passive*, and 3 from *passive-active*), 2 had response times under 2 seconds on the post or delayed test (1 from *passive*, 1 from *passive-active*). The analyses were carried out with 75 participants who have completed all phases of the study.

PALMs. In all three PALMs, participants were asked to map (and/or to study the correct mappings) between graphs and equations of Sine and Exponential functions. The AlgGeo PALMs focused on the following transformations of Sine and (natural) Exponential functions from the canonical functions $y = sin(x)$ and $y = e^x$ or $exp(x)$.

1) X-shifting: e.g., $y = sin(x + 4)$, $y = exp(x - 4)$

2) Y-shifting: e.g., $y = sin(x) + 4$, $y = exp(x) - 4$

3) X-scaling: e.g., $y = sin(x / 4)$, $y = exp(4x)$

4) Y-scaling: e.g., $y = 4sin(x)$, $y = exp(x) / 4$

These four transformations make up the four categories for training in each function family, making a total of 8 categories to be trained. Each of the transformation had 2 subtypes to account for the direction of the transformation. For example, the *x*-shifting category contains 2 subcategories: *x*-shifting to the left (e.g., $y = sin(x + 4)$) and *x*-shifting to the right (e.g., $y = sin(x – 4)$). Thus, each function family had 8 subcategories of transformation. There were 4-9 unique instances for each transformation subtypes, making a total of 119 functions used in the training[5].

In the *active* classification condition, participants practiced making active classifications with short mapping trials. On each trial, they were presented with a graph of a single function in blue and four equations, and were asked to pick from four equations the one that best matches the given graph. When the target was a Sine graph, all answer options contained only Sine functions, similarly for when the target was an Exponential function. On each trial, the module randomly selected 3 distractors from among the set of 7 possible alternatives (these together with the target represented 8 total subcategories for each function family), so that the second subtype of each transformation was one of the seven-distractor alternatives. All of the answer choices

[5] The compression subtypes included only 4 instances. This was because when the graphs within the compression subtypes were similarly scaled, they were very difficult to tell apart (e.g., $y = sin(x)/8$ and $y = sin(x)/9$). In later studies, we corrected for this issue by rescaling the graphs and added them back into the stimuli pool.

always involved the same quantities (e.g., if the target involves a shift or scale by 2 units, all distractor choices involve a transformation by 2 units). *Figure 2.1a* shows a sample *active* trial.

Accuracy and speed were continually tracked. After each response, there was sound and visual feedback to report whether the response was correct, and the speed of the response when correct. At feedback, the given graph was replaced with its contrastive version, in which the basic function ($y = sin(x)$ or $y = exp(x)$, as applicable) was added to the graph as a dotted gray line. *Figure 2.1b* shows a sample feedback screen. Each trial timed out after 30 seconds; at that point the screen cleared and participants were prompted to click Next to see the next trial. The trial feedback screen was not timed, so participants had an unlimited amount of time to view each trial feedback.

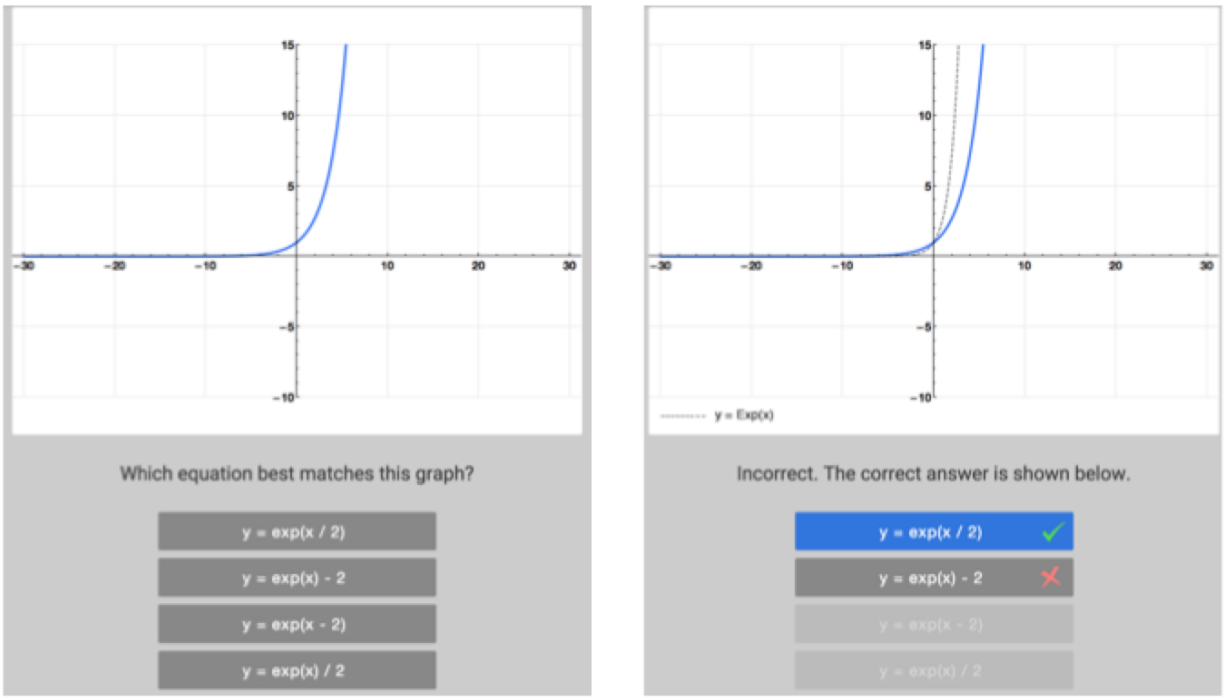


Figure 2.1. (a) Sample *active* trial and (b) its feedback.

A block feedback appeared after every 12 trials to provide the average block-by-block accuracy and RTc performance. This was an opportunity to take a break should the participant wanted it. When participants reached the preset learning criteria for each category (3 out of the last 3 trials of a category correct and each under 15 seconds), they gained mastery levels toward the completion of the module. This mastery feedback also appeared at block feedback as percentage completed to allow participants to view their cumulative progress in the PALMs. The module ended when participants complete all mastery levels.

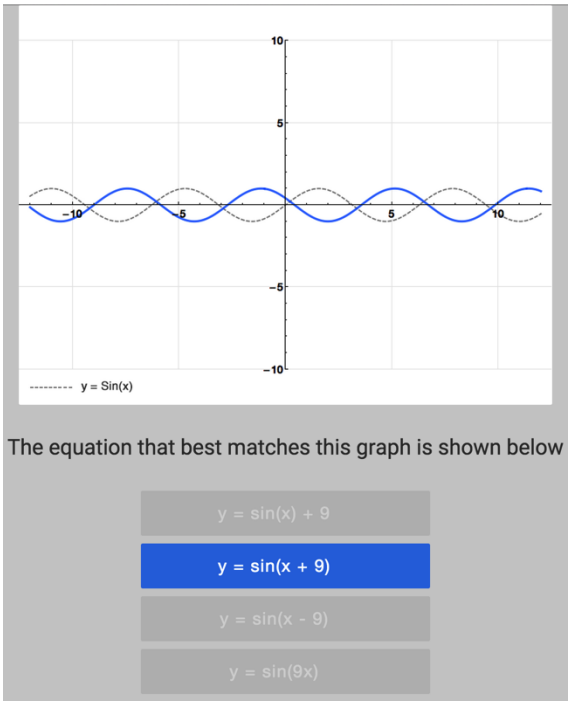


Figure 2.2. Sample *passive* trial.

In the *passive* training condition, on each trial a contrastive graph and its equation were shown as static images on the screen. This was the same display as the trial feedback screen (after a correct answer) as seen in the *active* condition, but here participants were asked to pay attention to figure out how to translate between the representations shown. *Figures 2.2* shows a sample *passive* trial. There was a sound to signal the start of a trial. The duration of each trial

was 14 seconds, the average amount of time it took for a pilot group of 15 *active* participants to make their choices and viewed the feedback. Participants were not asked to make a choice on each trial and thus did not receive feedback. After 14 seconds, the screen cleared, and participants were asked to click Next to continue on to the next trial. There were 171 trials in the *passive* training. This was the average number of trials it took for *active* participants to complete the training module. We also controlled for the difficulty of the categories by matching the proportion of trials from each category to the proportions seen in the *active* condition.

In the *passive-active* training condition, participants first received a set of 16 passive trials (2 examples from each category, 1 per subcategory), similar to those seen in the *passive* training condition, before participating in *active* classification trials. The *passive* trials were randomly presented, each for 14 seconds, in the same way as in the *passive* condition. In the *active* portion of the training, this group received the same feedback and completion criteria as those in the *active* classification condition. The duration of the training for those in the *active* classification condition and those in the *passive-active* condition depended on their performance during the modules.

Assessments

Three different versions of the assessment were given to participants in counterbalanced order as pretest, posttest and delayed test. The full list of assessment items is in the *Appendix E*. We modeled the question types after those used in Silva & Kellman (1999). Each question was presented in a similar multiple-choice format used in the training, with a graph and four equations as answer choices. There were 28 questions, divided into 4 types:

1) 8 Trained Items (TI). These were 4 Sine items and 4 Exponential items seen in the module. All transformation subtypes were represented.

2) 8 Trained Functions, Novel Items (TF/NI): These were new instances of trained transformations (4 Sine, 4 Exponential) involving different quantities than the ones seen in the training (i.e., the training used quantities from 2-10, and TF/NI items involved quantities 20, 30, and 40; e.g., $y = sin(20x)$). All transformation subtypes were represented.
3) 8 Untrained Functions (UF): These involved 4 Cosine and 4 Logarithmic functions, one for each trained transformation subtypes. All transformation subtypes were represented.
4) 4 Combination Functions (CF): These consisted of more complex combination functions, 2 of which involved a combination of Sine and Exponential functions (e.g., $y = sin(x) + exp(x)$), and the other 2 were combinations of transformations from one trained function family (e.g., $y = 3sin(x + 3)$).

Practice Questions. Each assessment (pretest, posttest, delayed test) was preceded by two practice questions to familiarize participants with the question format[6]. The practice questions (one Sine, one Exponential) were identical across all three versions of the assessment. These were designed to look like the rest of the assessment questions. Participants were not told that these were practice questions, and there were no feedback after each question, thus the assessment questions seamlessly followed suit.

Survey. To maximize the validity of our accuracy and response time measures, we asked participants to report honestly at the surveys their levels of engagement during the experiment, enjoyment of the training experience, whether they have sought outside help, what other activities they engaged in while participating in the experiment, whether they took breaks and what they did during the breaks, and of course, whether they had technical difficulties. The survey also contained questions about demographic information, their prior knowledge of the

[6] A pilot study showed that participants tended to timeout on the first question on the assessment.

materials, the amount of sleep they had the night before, a 4-item scale on their feelings of math anxiety toward math adapted from the Student Beliefs about Mathematics Survey (Kaya, 2008), the same four theory of intelligence items from Experiment 1, and some questions regarding their metacognitive judgments of their learning and prediction of their memory a week later. The full list of questions is in the *Appendix B.2*. The sequence of questions was the same for all participants.

Procedure

The experiment was web-based, and participants logged in from their own devices. They were instructed not take notes and to not consult outside resources nor to ask anyone for help at anytime during the study. All participants were told that the purpose of the experiment was to assess the effectiveness of this particular training program, and that they would be asked to engage in a training module to learn to map among Sine and Exponential graphs and equations, and that they would be tested on them afterward.

Figure 2.3 shows the procedure of the study. To complete the experiment, participants must qualify from the pretest – those who scored higher than 45% correctly on the pretest (averaged from all 28 trials) were deemed ineligible for the study. No feedback was provided after each pretest question. After the pretest, if they were eligible, participants had a 6-hour window to complete the training, the immediate posttest and the survey. We used a condition-balancing algorithm to assign eligible participants into three training conditions based on their pretest accuracy to ensure similar overall pretest accuracy and equal distribution of participants across conditions[7]. They were given 6 hours to complete the training, take the immediate posttest, and fill out the survey. A week after the immediate posttest, participants took the

[7] This assignment algorithm was not available when the ECG experiments were conducted.

delayed test and completed another survey. They received instructions for the delayed test via email the day before.

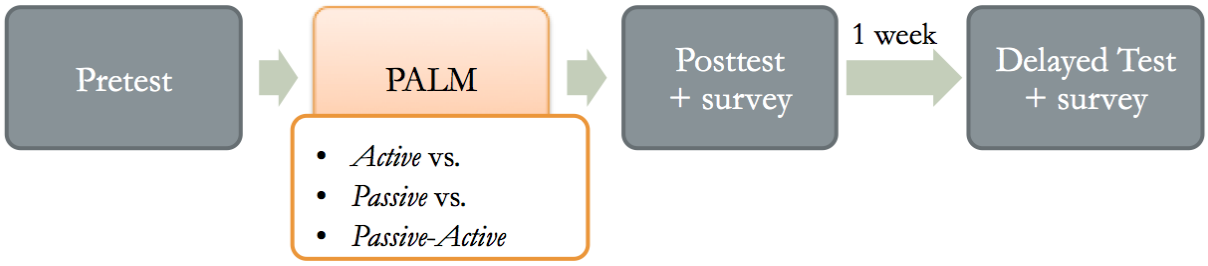


Figure 2.3. Procedure of Experiment 2

Overview of Analyses and Expected Results

We collected posttest data only from participants who have completed the training modules, thus we do not report results from everyone who had attempted the module and did not finish and from participants who did not return for the delayed test. We collected data until we acquired equal number of participants in each condition, after excluding those who self-reported to have sought help anytime during the one-week duration of the study, and from those who reported to have experienced distractions and technical difficulties.

We report these results for all assessment item types together and separately for each item type. Similar to the previous experiment, we report analyses from a 3 phase (pretest, posttest, delayed test) x 3 condition (*active, passive, passive-active*) mixed ANOVAs on raw accuracy and fluent accuracy. While we successfully equated the overall pretest accuracies across groups, they were slightly different on some assessment trial types. Whenever appropriate in the following analyses, we also conducted a 2 phase (pre-post, pre-delayed test) x 3 condition (*active, passive, passive-active*) mixed ANCOVA on efficiency, accuracy gain, and fluency gain. All assumptions were met.

Since improvement in information extraction can theoretically produce both near and far transfer to cases of familiar and more complex functions, we expected to see overall posttest gains from all training conditions. In terms of condition differences, we expected to replicate Experiment 1's finding that passive exposure to multiple varied instances can improve participants' ability to discover relevant relations and enhance learning and retention when followed by active classification training. We expected the *passive-active* condition to yield better efficiency scores than the *active* condition, and to outperform the *active* and *passive* conditions on transfer and retention.

RESULTS

Efficiency

Efficiency by Trial

Figure 2.4a displays the efficiency by total number of trials completed. A 2 phase x 3 condition ANCOVA with pretest accuracy as the covariate confirmed a main effect of condition, $F(2,71) = 4.35$, $p = .02$, $\eta^2_p = .11$, and no phase x condition interaction, $F(2,71) = .11$, $p > .20$. The *passive-active* condition ($M = .0017$, $SD = .0015$) produced higher overall trial efficiency than the *passive* condition with a medium effect size ($M = .0008$, $SD = .0007$), $t(48) = 2.87$, $p = .006$, $d = .76$. This difference was reliable at both phases (pre-post: $t(48) = 2.71$, $p = .009$, $d = .73$, and pre-delayed, $t(48) = 2.87$, $p = .006$, $d = .77$). *Passive-active* also had numerically higher overall trial efficiency than *active* ($M = .0012$, $SD = .0012$), but this difference was not statistically significant, $t(48) = 1.50$, $p = .14$. However, planned comparisons showed that even though *passive-active* did not differ from *active* at pre-post trial efficiency, $t(48) = 1.09$, $p = .28$, *passive-active* had marginally higher pre-delayed trial efficiency than *active* with a medium effect size, $t(48) = 1.69$, $p = .097$, $d = .50$. There was a reliable difference between the *active* and

passive groups on overall trial efficiency, $t(48) = 1.31$, $p = .20$. This held at both pre-post, $t(48) = 1.19$, $p = .24$, and pre-delayed, $t(48) = 1.06$, $p = .29$, phases.

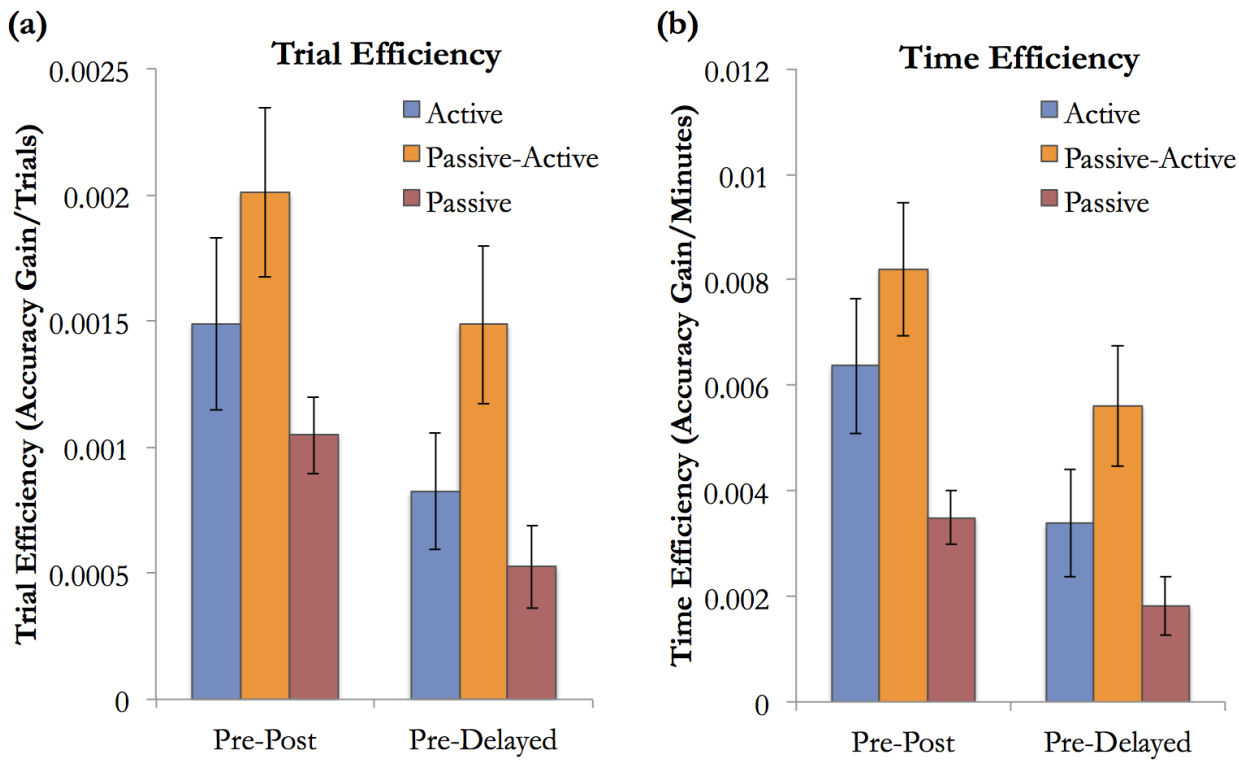


Figure 2.4. Efficiency (a) by trial and (b) by time. Error bars are ± 1 standard error.

There was no significant main effect of phase, $F(1,72) = 17.85$, $p < .001$, $\eta^2_p = .20$, confirming that after controlling for variations at pretest accuracy, pre-post trial efficiency was not reliably different from pre-delayed trial efficiency. There was no phase x pretest interaction, $F(1, 71) < 1$, $p > .20$, and there was a marginal main effect of pretest accuracy, $F(1,71) = 3.02$, $p = .09$, $\eta^2_p = .04$, but the correlations between pretest accuracy and pre-post trial efficiency, $r(75) = -.14$, $p = .22$, and with pre-delayed trial efficiency, $r(75) = -.19$, $p = .11$, were not statistically reliable.

By Assessment Item Types

The patterns of condition differences on trial efficiencies varied slightly by item type. On trained items (TI), both the *passive-active* (M = .0022, SD = .0023) and the *active* (M = .0017, SD = .0021) conditions produced higher overall TI efficiency than the *passive* (M = .0007, SD = .0013) condition with medium effect sizes, $t(48) = 2.71$, $p = .009$, $d = .80$, and $t(48) = 1.98$, $p = .05$, $d = .57$, respectively. There were no reliable differences between the *passive-active* and the *active* conditions on the overall TI efficiency, $t(48) = .75$, $p = .46$, both at pre-post, $t(48) = .23$, $p = .82$, and at pre-delayed, $t(48) = 1.07$, $p = .29$, phases.

On novel items of trained functions (TF/NI), the *passive-active* condition (M = .0021, SD = .0020) surpassed the *passive* condition on overall trial efficiency (M = .0006, SD = .0011) with a large effect size, $t(48) = 3.28$, $p = .002$, $d = .95$, and also had marginally higher efficiency than the *active* condition (M = .0011, SD = .0020) with a medium effect size, $t(48) = 1.81$, $p = .08$, $d = .53$. The difference between *passive-active* and *active* was not statistically significant at pre-post, $t(48) = .97$, $p = .34$, but it was statistically reliable at pre-delayed phase with a medium effect size (.0019 vs. .0005, respectively), $t(48) = 2.26$, $p = .03$, $d = .65$.

There were no condition differences on untrained functions (UF) and combination functions (CF), $t(48) < 1$, p's $> .20$. More details of these results are in *Appendix D.2*.

Efficiency by time

Figure 2.4b displays the efficiency by total time invested. Unlike trial efficiency, both the *passive-active* condition (M = .007, SD = .005) and the *active* condition (M = .005, SD = .005) produced higher overall time efficiency than the *passive* condition (M = .003, SD = .002) with medium and large effect sizes, $t(48) = 3.57$, $p = .001$, $d = .99$, and $t(48) = 2.06$, $p = .04$, $d = .58$, respectively. There was no reliable difference between the *passive-active* and *active* conditions

on overall time efficiency, $t(48) = 1.39$, $p = .17$, at pre-post, $t(48) = 1.02$, $p = .31$, nor at pre-delayed, $t(48) = 1.44$, $p = .16$, phases.

By Assessment Item Type

On TI time efficiency, *passive-active (*$M = .0083$, $SD = .0089$) did better than *passive* ($M = .0026$, $SD = .0043$) with a large effect size, $t(48) = 2.88$,$p = .006$, $d = .82$, but did not differ from *active* ($M = .0083$, $SD = .0097$), $t(47) < 1$, $p > .20$. *Passive-active* was not different from *active* at both pre-post and pre-delayed phases, $t(47) < 1$, p's $> .20$. The *active* group had higher overall TI time efficiency than *passive*, $t(48) = 2.68$, $p = .01$, $d = .76$, which was reliable at pre-post, $t(48) = 3.35$, $p = .002$, $d = .95$, but not at pre-delayed phase, $t(48) = 1.46$, $p = .15$, $d = .42$.

On TF/NI, the *passive-active* group had marginally higher efficiency than the *active* group with a medium effect size (.008 vs. .005), $t(48) = 1.74$, $p = .09$, $d = .50$. Their difference was not reliable at pre-post, $t(48) < 1$, $p > .20$, but was reliable at pre-delayed phase, $t(48) = 2.09$, $p = .04$, $d = .58$. The difference between the *active* group and *passive* group on overall TF/NI time efficiency was marginally significant and with a small effect size (.0046 vs. .0021, respectively), $t(48) = 1.68$, $p < .10$, $d = .48$. This was significant at pre-post with a large effect size (.0075 vs. .0027), $t(48) = 3.08$, $p = .003$, $d = .88$, but not at pre-delayed phase (.0018 vs. .0016), $t(48) < 1$, $p > .20$.

There were no condition differences on trial efficiencies when measured with UF and CF, $t(48) < 1$, p's $> .20$. Additional details of these analyses may be found in *Appendix D.2.*

Accuracy

All Items

Figure 2.5a shows the mean accuracy on all assessment items. As we expected, across all conditions, PALM training led to strong overall improvement and retention. A 3 phase x 3

condition ANOVA confirmed a main effect of phase, $F(2,144) = 71.00$, $p < .001$, $\eta^2_p = .50$. Across all conditions, there were reliable improvements from pretest ($M = .32$, $SD = .07$) to immediate posttest with very large effect sizes ($M = .51$, $SD = .15$), $t(74) = 11.38$, $p < .001$, $d = 1.62$. There was a reliable drop from immediate posttest to delayed test ($M = .43$, $SD = .15$), $t(74) = 4.97$, $p < .001$, $d = .53$, but a substantial amount was retained, relative to pretest, a week later, $t(74) = 7.09$, $p < .001$, $d = .94$. The improvement from pretest to delayed test held for all item types (see *Appendix D.2* for more details of these analyses).

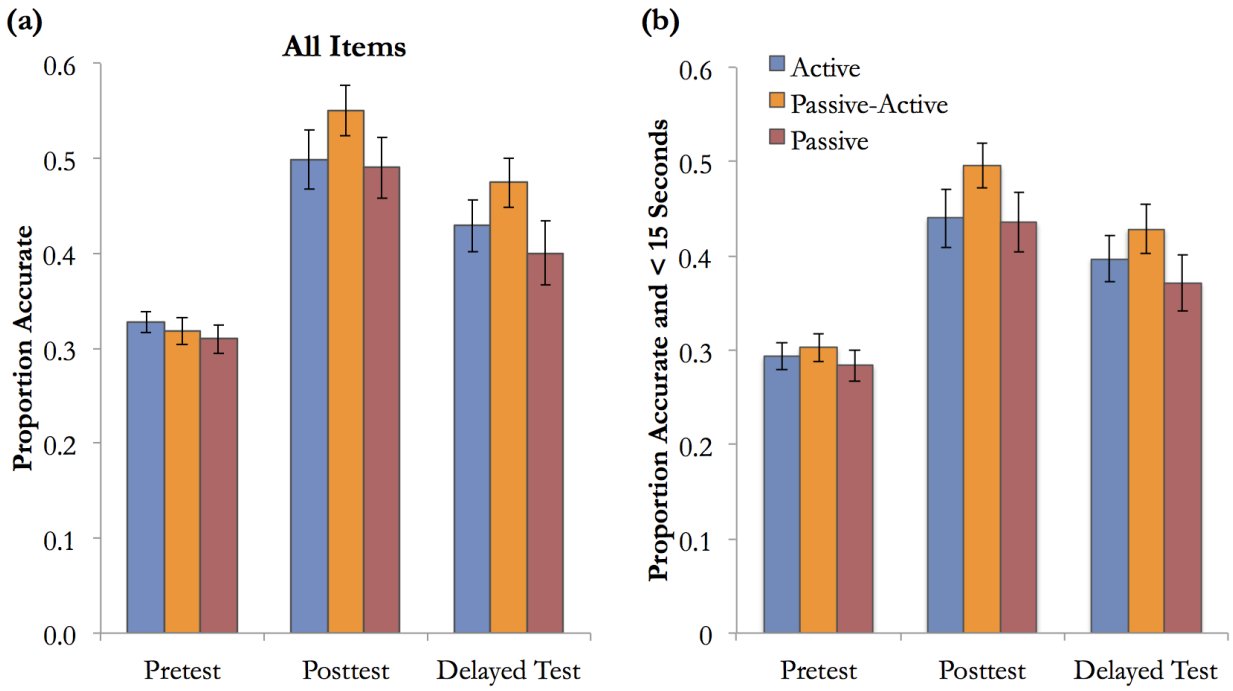


Figure 2.5. Mean (a) accuracy and (b) fluent accuracy on all assessment items. Error bars are ± 1 standard error.

There was no reliable overall main effect of condition, $F(2,72) = 1.51$, $p = .23$, $\eta^2_p = .04$, and no phase x condition interaction, $F(4,144) = 1.01$, $p = .40$, $\eta^2_p = .03$. Planned comparisons showed that *passive-active* had marginally higher accuracy at delayed test than *passive* with a medium effect size, $t(48) = 1.75$, $p = .09$, $d = .50$, but there was no reliable difference at

immediate posttest, $t(48) = 1.44$, $p = .16$. The *passive-active* group had numerically higher accuracy than *active*, but these differences were not reliable at immediate posttest nor at delayed test, $t(48) < 1.3$, p's $> .20$. There were no reliable differences between *active* and *passive* at any phases, $t(48) < 1$, $p > .20$.

By Assessment Item Type

Condition differences were more apparent on trained items (TI) and trained functions, novel items (TF/NI). *Figure 2.6a* displays the mean accuracy on TI, and *Figure 2.7a* displays the mean accuracy on TF/NI. All conditions produced large and enduring improvements on TI and TF/NI from pretest to immediate posttest and from pretest to delayed test, with some reliable forgetting between immediate posttest and delayed test, p's $< .01$.

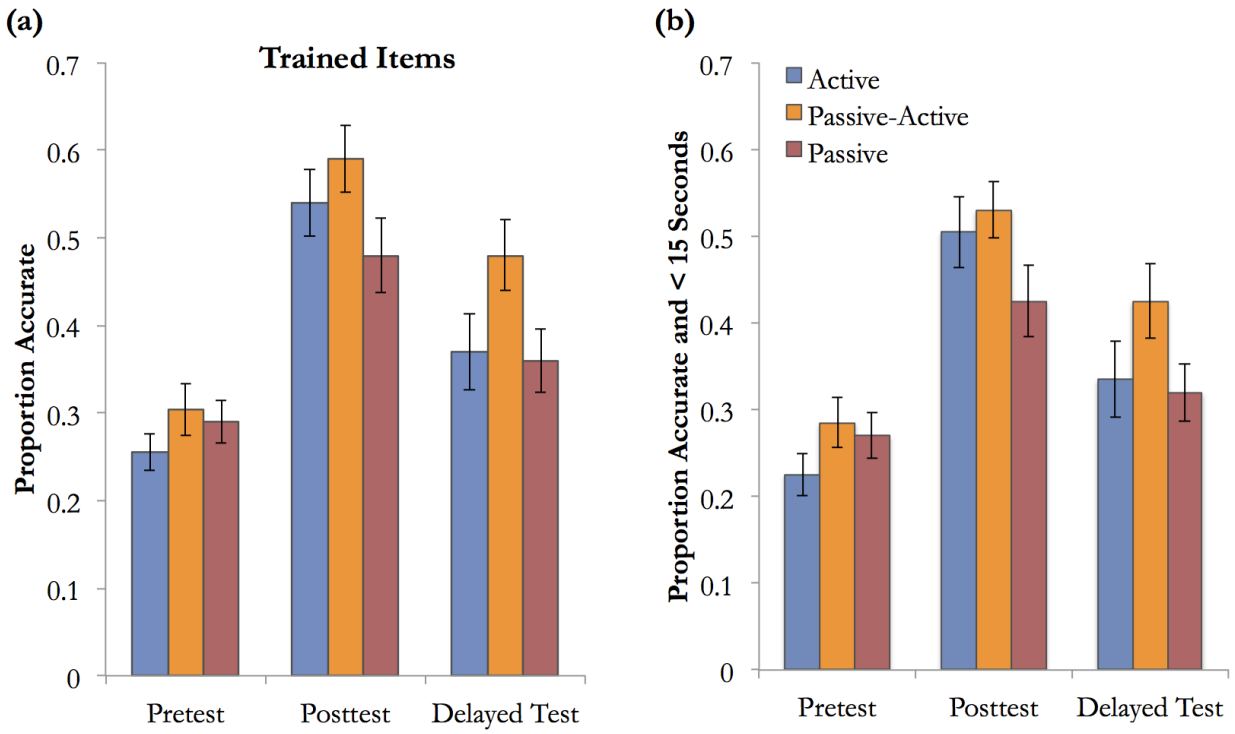


Figure 2.6. Mean (a) accuracy and (b) fluent accuracy on Trained Items (TI).

Error bars are ± 1 standard error.

For trained items (TI), the *passive-active* group did better than the *active* group on overall accuracy (.46 vs. .39), $t(48) = 2.12$, $p = .04$, $d = .61$. The difference was not statistically reliable at immediate posttest (.59 vs. .54, $t(48) < 1$, $p > .20$), but it was marginally reliable at delayed test (.48 vs .37), $t(48) = 1.86$, $p = .07$, $d = .53$. The *passive-active* group also did better than the *passive* group on overall accuracy with a medium effect size (.46 vs. .38), $t(48) = 2.25$, $p = .03$, $d = .64$. This difference was marginal at immediate posttest (.59 vs. .48), $t(48) = 1.92$, $p = .06$, $d = .54$, and significant at delayed test (.48 vs. .36), $t(48) = 2.22$, $p = .03$, $d = .63$. The *active* and *passive* groups did not differ at either posttests, $t(48) < 1.1$, $p > .20$.

In terms of TI accuracy gains, *passive-active* had numerically higher overall accuracy gain than *passive,* but the difference was not statistically significant overall, $t(48) = 1.66$, $p = .10$, $d = .47$, on pre-post gain, $t(48) = 1.47$, $p = .15$, nor on pre-delayed test gain, $t(48) = 1.60$, $p = .12$. *Passive-active* did not differ from *active* overall, on pre-post gain, nor on pre-delayed gain, $t(48) < 1$, p's $> .20$. Similarly, *active* did not differ from *passive* overall, on pre-post gain, nor on pre-delayed gain, $t(48) < 1.5$, p's $> .10$. Participants had similar accuracies for Sine and Exponential TI. The same patterns of result were found for both function families, but the condition differences were more pronounced with Exponential TI. *Appendix D.2* contains more details of these analyses.

For novel items of trained functions (TF/NI), there were no reliable differences among conditions at pretest, posttest, nor delayed test, $t(48) < 1.7$, $p > .10$, but notable condition differences were found in terms of TF/NI accuracy gain. *Passive-active* had higher overall gain than *passive* with a medium effect size (.23 vs. .11), $t(48) = 2.42$, $p = .02$, $d = .68$, as seen from pre to posttest, $t(48) = 2.11$, $p = .04$, $d = .60$ and from pre to delayed test, $t(48) = 2.04$, $p = .047$, $d = .58$. *Passive-active* also had numerically higher accuracy gain on TF/NI than *active,* but only

marginally significantly so on pre-delayed gain with a medium effect size, $t(48) = 1.84$, $p = .07$, $d = .52$. Their difference on the overall gain and on pre-post gain was not reliable, $t(48) = 1.45$, $p = .15$ overall, and $t(48) < 1$, $p > .20$ on pre-post gain. Similarly, *active* did not differ from *passive* overall, $t(48) < 1$, $p > .20$, on pre-post gain, $t(48) = 1.34$, $p = .19$, nor on pre-delayed gain, $t(48) < 1$, $p > .20$. Interestingly, these condition differences were driven by the same differences on Exponential TF/NI items. There were no condition differences on Sine TF/NI.

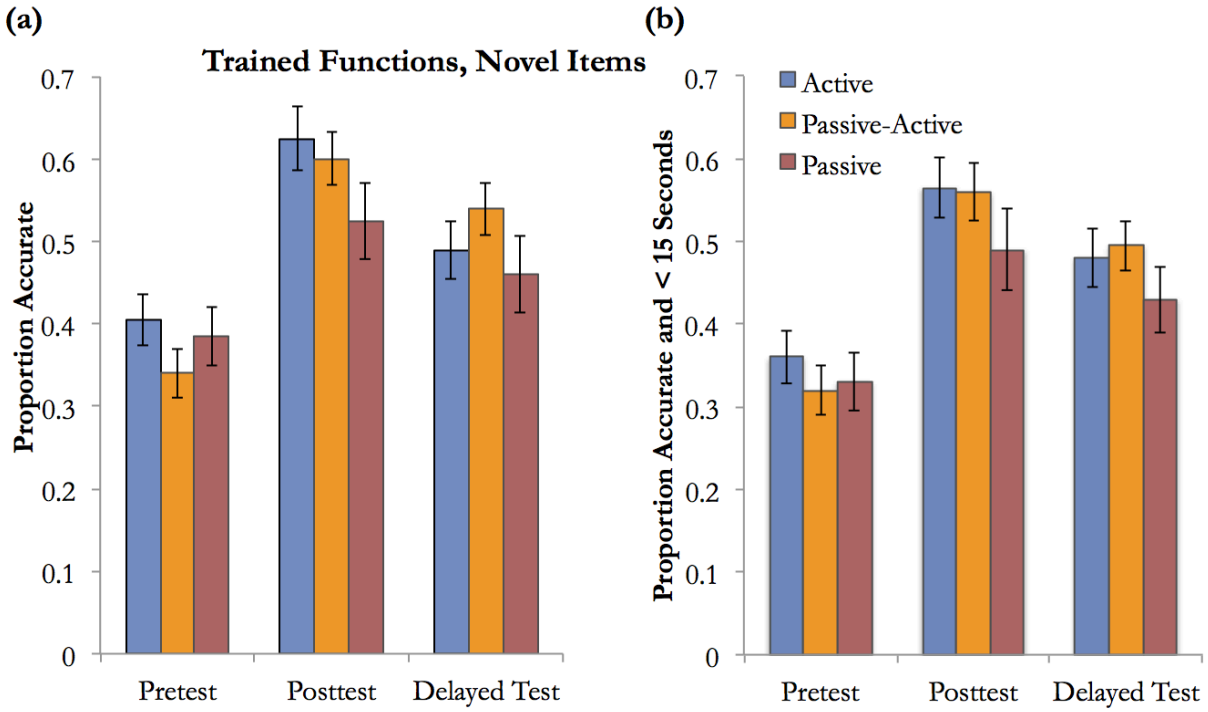


Figure 2.7. Mean (a) accuracy and (b) fluency on Trained Function, Novel Items (TF/NI). Error bars are ± 1 standard error.

The only notable condition differences on untrained functions (UF) were that the *passive-active* condition had marginally higher accuracy than the *active* condition at immediate posttest (.51 vs. .40), $t(48) = 1.98$, $p = .05$, $d = .56$, but this difference did not appear at delayed test (.47 vs. .42), $t(48) < 1$, $p > .20$, nor in terms of UF accuracy gains, $t(48) < 1.2$, $p > .20$. Also, the

passive condition had marginally higher pre-post UF accuracy gain than *active* (.22 vs. .10), $t(48) = 1.82$, $p = .08$, $d = .43$, but not on pre-delayed UF gain, $t(48) < 1$, $p > .20$. There were no reliable condition differences on combination functions (CF), $t(48) < 1.6$, $p > .16$.

Interesting, all three group showed reliable pre-delayed test improvements (but not pre-post) on UF with medium effect sizes, p's < .02, d = .57 to .61. Pre-post learning gains on CF were reliable for the *passive-active* and *passive* groups, but not pre-delayed gains, p's > .20. *Appendix D.2* contains more details of these analyses.

Fluency

All Items

Figure 2.5b shows the fluent accuracy by condition. Like raw accuracy, overall there were strong improvements from pretest to immediate posttest that sustained to delayed test, and a small drop in performance between immediate posttest and delayed test, p's < .001. Unlike raw accuracy, there were no condition differences on overall fluent accuracy.

By Assessment Item Types

Fluent accuracy on trained items (TI; *Figure 2.6b*) show similar pattern as raw accuracy on TI. The advantage of *passive-active* over *active* was only marginally significant and had a small effect size, $t(48) = 1.71$, $p = .09$, $d = .42$. This difference was not reliable at immediate posttest, $t(48) = .48$, $p > .20$, nor at delayed test, $t(48) = 1.47$, $p = .15$. *Passive-active* was better than *passive* overall, $t(48) = 2.14$, $p = .04$, $d = .61$. This difference was marginally significant but with medium effect sizes at immediate posttest, $t(48) = 1.99$, $p = .05$, $d = .56$, and at delayed test, $t(48) = 1.94$, $p = .06$, $d = .55$.

On trained functions, novel items (TF/NI; *Figure 2.7b*), unlike raw accuracy, there were no reliable condition differences, *p*'s > .10.

Interestingly, and unlike raw accuracy, there were marginally reliable condition differences on overall untrained functions (UF) fluent accuracy, $F(2,72) = 2.44$, $p = .09$, $\eta^2_p = .06$. The *passive-active* group ($M = .40$, $SD = .08$) had higher overall UF score than the *active* group with a medium effect size, $t(48) = 2.26$, $p = .03$, $d = .66$, and marginally higher than the *passive* group with a small effect size, $t(48) = 1.75$, $p = .09$, $d = .49$. There was no difference between the *active* ($M = .34$, $SD = .10$) and *passive* groups ($M = .35$, $SD = .12$), $p > .20$.

There were no reliable condition differences on combination function (CF) items, *p*'s > .20.

Progression of Learning

The *active* and *passive-active* conditions did not differ on accuracy during in the training, but they differed on overall fluent accuracy, $t(48) = 2.01$, $p = .05$, $d = .59$. *Table 2.1* shows the mean proportion of accurate and fluent accurate scores on *active* trials from the *active* and *passive-active* trainings.

Condition	Trials Completed	Minutes on Training	Training Accuracy	Training Fluent Accuracy
Active	171.2 (16.2)	31.5 (2.53)	.41 (.02)	.38 (.02)
Passive	171.4 (1.2)	51.9 (2.67)	--	--
Passive-Active	150.0 (13.6)	33.1 (2.67)	.45 (.02)	.43 (.02)

Table 2.1. Training means by condition. Standard errors are in parentheses.

By Quartile

Figure 2.8 shows the mean accuracy and fluent accuracy by quartiles and by blocks. In terms of accuracy, the condition differences did not reach statistical significance in a 4 quartile x

2 condition (*active* and *passive-active*) ANOVA, $F(1,48) = 2.13$, $p = .15$, $\eta^2_p = .04$. There was a main effect of phase, $F(3,144) = 21.34$, $p < .001$, $\eta^2_p = .31$, reflecting a steady increase in accuracies during the training modules. There were similar mean accuracies during the first two quartiles, $p > .10$. This was likely a reflection of the adaptive retirement feature of the modules. As soon as the learner reached learning criterion for a particular category, items from that category dropped out from the module to focus the remaining training time on the remaining categories. This also explained why accuracy on the module did not reach 100%. Accuracy improved from the second to the 3rd quartile ($M = .38$, $SD = .11$ to $M = .45$, $SD = .15$), $t(49) = 3.40$, $p < .01$, $d = .53$, and from the third to the fourth quartile ($M = .53$, $SD = .16$), $t(49) = 3.66$, $p < .01$, $d = .52$. There was no quartile x condition interaction, $F(3,144) = .63$, $p > .10$, $\eta^2_p = .01$.

Fluent accuracy showed the same pattern of accuracy improvements, except that there was also a main effect of condition, $F(1,48) = 4.10$, $p < .05$, $\eta^2_p = .08$. The *passive-active* condition ($M = .43$, $SD = .08$) tended to have higher fluent accuracy throughout the training than the *active* condition ($M = .38$, $SD = .08$).

By Blocks

The advantage of the *passive-active* over *active* could be seen at block 2 (their first block in the active classification portion). This effect had a medium effect size in terms of accuracy, $t(48) = 2.59$, $p = .01$, $d = .73$, and a large effect size in terms of fluent accuracy, $t(48) = 2.83$, $p = .01$, $d = .80$.

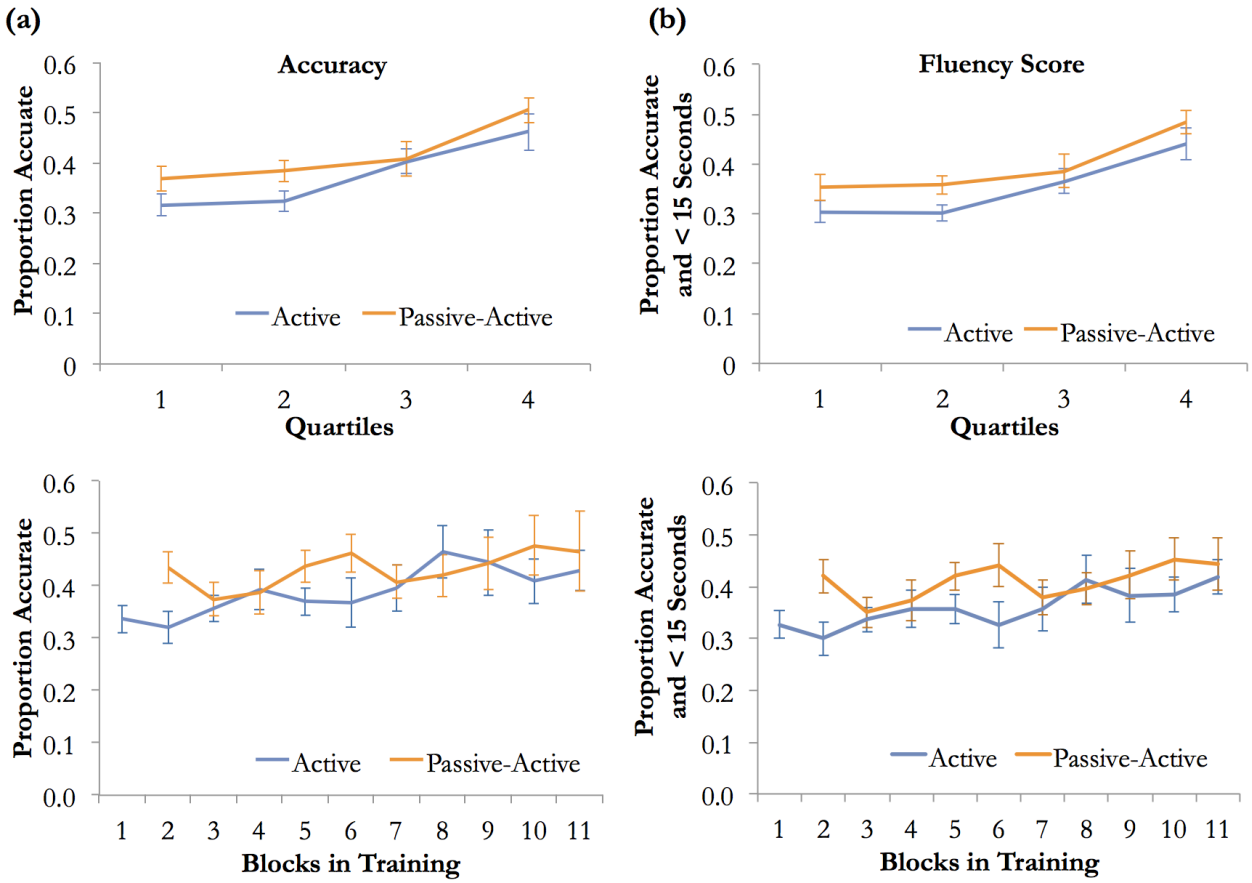


Figure 2.8. Mean (a) accuracy and (b) fluent accuracy during the training, by quartiles (top) and by blocks (bottom). Error bars are ± 1 standard error.

As a result, the *passive-active* group took about 20 fewer trials (including both *passive* and *active* trials) to reach learning criteria (M = 150, SD = 67.94) than the *active* group (M = 171, SD = 81.02). However, this difference did not reach statistical significance, $t(48) = 1.00$, $p > .10$. These two groups also did not differ on the total training duration, $p > .10$. The *passive* training, however, took almost 20 minutes longer in the training than the *active* condition. Each *passive* trial only lasted 14 seconds, but participants in the *passive* training seemed to take a few seconds longer on each trial on average before continuing on to the next trial.

Self-Reported Measures

There were no condition differences on any of the self-reported ratings. Participants generally were quite engaged and motivated (M = 5.32, SD = .93), and found the training modules to be fairly enjoyable (M = 3.96, SD = 1.49), and helpful (M = 4.11, SD = 1.30). All questions were on a 1-6 scale, with 1 being "Not at all", and 6 being "Very much". They self-reported not knowing much about Sine and Exponential functions prior to the study (M = 2.05, SD = 1.00) but they were conservative in their rating of their knowledge immediately after the study (M = 3.32, SD = 1.22) and a week later (M = 2.55, SD = 1.12). They also did not have strong confidence on their performance on the delayed test (M = 3.00, SD = 1.19).

DISCUSSION

All training conditions produced durable improvements in accuracy and fluency on TI and TF/NI, and UF at a delay. Overall, the *passive-active* condition produced higher training efficiency measured both by trials and by time invested than the *passive* condition. *Passive-active* had higher efficiencies than the *active* training only when they were calculated with TF/NI accuracy gain. The *active* condition was also more efficient than the *passive* condition on trial efficiency when calculated with TI accuracy gain, and on time efficiency. Overall there were no differences in accuracy between these two conditions, but those in the *passive* training tended to spend slightly longer on each trial before moving on to the next. It should be noted that there the actual viewing duration per trial were similar between *active* and *passive* conditions. Each *passive* trial timed out after 14 seconds, at which point the graph was cleared from the screen. The benefit of *passive-active* was particularly robust on overall TI accuracy, and on TF/NI accuracy gain (though only at the pre-delayed gain when compared to the *active* condition). The *passive-active* group was not reliably more accurate but had greater fluent accuracy than the

active condition. Effect sizes for condition differences ranged from medium to large effect sizes.

There were no condition differences on the enjoyability and helpfulness ratings. One possible reason was that for our Amazon Mechanical Turk participants, this study was relatively more interesting despite being longer and more cognitively demanding compared to the many other surveys they were used to. When asked to leave comments about the study, participants across conditions generally thought the study was "difficult but fun and interesting".

GENERAL DISCUSSION

The benefits of *passive-active* training generally held across two very different learning domains. In Experiment 1, having just two instances per category shown passively prior to undertaking the active classification task dramatically elevated performance during the training module and at assessments, particularly on transfer items never seen in the training. The advantage of *passive-active* over just *passive* training was robust across all measures. *Passive-active* training also led to better transfer performance than *active.* It was also more efficient than *active* when we assumed that pretest variations were due to chance. This advantage of *passive-active* training over just *active* classification and just *passive* exposure was most apparent at the one-week delay. *Active* also proved to be more efficient than *passive* when we corrected for pretest variations, and to be more effective than *passive* for transfer to new instances at a delay.

In Experiment 2, we replicated the benefit of *passive-active* and *active* for enhancing efficiencies over the *passive* condition. The difference in overall efficiencies between *passive-active* and *active* was not reliable, except when considered for TF/NI accuracy gains. This is still noteworthy, as the ability to transfer what was learned to new instances of the learned categories is an important goal for training. In some cases, *passive-active* showed reliably better learning than *active*, such as on Exponential TI and TF/NI accuracy at a delay.

Despite some nuances in our findings, the overall pattern supported the benefit of *passive-active* over *passive*, and in most cases, *passive-active* was generally more effective and efficient than *active,* as *active* was more so than *passive* in some cases. We concluded that a brief passive study as a primer to active classification could be a potent method for improving accuracy and efficiency in adaptive perceptual learning.

Theoretical Implications

The advantages of *passive-active* learning have several possible explanations. Consistent with work on cognitive load and worked examples (e.g., Renkl et al., 2004), initial familiarization with category exemplars may allow relevant structure to be learned without imposing the additional task demands of active responding. Moreover, passive and active learning may complement each other in focusing attention on within-category similarities and between-category contrasts respectively (e.g., Carvalho & Goldstone, 2014). The initial passive study provides an opportunity for learners to understand the specific features or relations that define each category, which supported the discrimination process in active classification. Other specific advantages of passive exposure at the start of learning may include avoiding initial errors and persistence of incorrect guesses, as well as averting frustration that may arise in active learning from having to guess initially. Taken together, the combination of passive and active practice may enhance learners' ability to pick up on the relevant structures earlier on, so that the remaining practice could support improvements in fluency. As a result, *passive-active* training effectively produced changes in perceptual processes needed for durable learning. In both experiments, the benefits of *passive-active* were found mainly after a long retention interval, rather at immediate posttest. A signature perceptual learning effect is little decay over time (Kellman et al., 2009). This is consistent with other work showing little decay as a result of

perceptual learning (Kellman et al., 2009; Krasne et al., under review). There was very little drop after a delay for categories learned fluently. This result provides evidence of the benefits of response times in mastery criteria (Mettler, Massey, & Kellman, 2011). It is also similar to other studies showing that retention interval has been shown to moderate the benefit of learning techniques. For example, practice testing and spacing both have larger effects after longer versus shorter retention intervals (for reviews, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013).

Both Experiments 1 and 2 showed benefits of passive exposure alone for perceptual learning. Because of the nature of the two different learning domains, the design of the *passive* tasks differed. This may have led to differences in what was learned from the *passive* trials. In Experiment 1, aside from the category label, each diagnostic feature was described and marked on the ECG. The mark-ups were provided, but the features may not have been apparent immediately to the learners, as the benefit of *passive-active* over *active* on the active classification task did not appear until block 4 in the training. This suggested that only after classification practice with multiple instances that participants began to recognize the feature variations. In Experiment 2, on the other hand, there were no category labels nor descriptions of the category (such as "the graph belongs to the "*x*-shifting" category because the function was shifted to the right by 4 units"). Instead, participants were shown the matching equation (e.g., $y = sin(x - 4)$). This passive task produced an immediate advantage for *passive-active* at block 2. One possibility is the passive trials in this case involved more inferential learning during which the learners had to infer for themselves the diagnostic patterns, which effectively enhanced their active classification learning later on.

In both cases, passive exposure proved to be effective to guide attention to the relevant

features and relations for effective transfer in some cases as active classification at delayed tests. Although the nominal task during learning in the *passive* condition and the active classification test were not identical (i.e., students in the *passive* group were explicitly told which heart pattern was illustrated during learning versus having to classify ECG patterns), the *passive* learning task afforded practice with seeing how the features could be mapped onto aspects of novel traces. These findings are consistent with prior work on category learning by observation, that when provided with the category label, learners have to infer for themselves the diagnostic features, and doing so can enhance the extraction of the particular relational structure, within-category correlations (e.g., Yamauchi & Markman, 2000; Chin-Parker & Ross, 2002; Levering & Kurtz, 2015) that are effective for supporting later classification tasks.

Active classification in most cases produced higher accuracy gain and efficiency than *passive* exposure. One may argue that this is in accordance to the transfer-appropriate-processing framework (Morris, Bransford, and Franks, 1977), such that memory is enhanced to the extent that the cognitive processes engaged during the learning overlap with those engaged during the assessment task. There is a possibility that the functional overlap between the learning task and the final test may have benefited the *active* and *passive-active* groups in ways other than enhancing the overlap of conceptual processes (e.g., although the particular items included in the novel classification task had not been seen earlier, learners in the *active* group may still have benefited from overall familiarity with a task that involved active classifications of examples). With that said, other literatures have shown that task overlap per se does not always confer benefits for subsequent performance. For example, Karpicke and Blunt (2011) demonstrated that retrieval practice was more effective than concept mapping practice for enhancing performance on a subsequent concept mapping test, and Paas and Van Merriënboer (1994) demonstrated that

studying worked examples of geometry problems was more effective than solving geometry problems for enhancing performance on a subsequent problem-solving test. In the current experiments, the advantage of *passive-active* over *active* was evident, despite participants in both conditions have reached the same learning criteria on the classification task, suggesting that what was learned was better extraction of relevant information from novel stimuli, rather than mere overlap of task requirements.

Practical Implications

This work has clear practical implications. Incorporating *passive-active* training is an easily implemented technique that is likely to improve learning technology. It also proved to be efficient and enjoyable for learners, especially when compared to the *passive* training in ECG interpretation training.

One striking finding from Experiment 1 was that the primer had very limited effect on learning. The *primer* was modeled after textbook instruction. It prepared undergraduates to benefit from the ECG PALMs, but it was clearly not sufficient for producing highly accurate or fluent interpretation of heart patterns (e.g., accuracy levels after the *primer* averaged around 30% (pretest accuracy in *Figure 1.5*). The common assumption of classroom instruction is if a lecture has been delivered clearly, or if an example or two has been worked in detail, then an attentive, earnest student should absorb the relevant concepts. Yet, many students year after year seem resistant to absorbing the basics of scientific or mathematical concepts, and even fewer can apply these productively in problem solving. Recent studies assessing the severity of ECG interpretation errors have reported that upwards of 33% of interpretations of medical professionals and residents contained errors of major importance in clinical settings (Mele, 2008; see also Krasne, Stevens, Kellman & Neimann, under review). These difficulties may seem

baffling from the perspective of conventional instruction, but they are understandable (and to be expected) from the perspective of perceptual learning. The emphasis in declarative knowledge, the stating of facts and concepts, ignoring how experts can classify problems but they often are unable to convey their insight in how they solved the problem verbally.

The learning domains tested here involved complex presentations with many varying information, some relevant and some irrelevant. A graph of a function, for example, contains crucial information in terms of the shape of a function, scaling, etc. These information may seem obvious to a teacher, but a novice does not intuit immediately what are the relevant and irrelevant features of the representation. In the case of ECG, consider a cardiologist instructing her students on how to distinguish between right bundle branch block and anterior STEMI by looking for the elevation in ST segments versus a broad QRS complex. This information can guide students to know what features to look for, but doesn't guarantee that students can *see* them when given a new ECG to consider. These perceptual features require many hours (or even months or years) of training to develop (Wood, Batt, Appelboam, Harris & Wilson, 2013; Salerno, Alguire, & Waxman, 2003; Mele, 2008), and the ability to fluently recognize them when they are relevant requires perceptual training by witnessing systematically varying cases (Goldstone, Landy, & Brunel, 2011; Kellman & Massey, 2013). The descriptions from our *primer* did not directly alter the internal workings of perceptual processes, but were nonetheless useful for focusing one's attention on different aspects of a tracing.

Perceptual learning processes advance when students makes rapid classifications of varying instances and receive feedback, especially following some passive exposure to the correct classifications. The present results confirm the importance of perceptual learning

interventions as a valuable complement to declarative and procedural components of instruction (Kellman & Massey, 2013).

Limitations and Future Directions

While our results suggest that participants retained a great amount at delayed test, they had also forgotten some of what they had learned after a week delay. Accuracy of classification at posttests alone, however, may not accurately reflect how much they remembered. It could be that participants could well differentiate between the ECG patterns but they had forgotten the label associated with each pattern. Future studies can incorporate a matching task at posttests to parse apart perceptual learning effects from memory of labels.

Renkl et al. (2000) suggested that a possible reason why the *passive-active* group learned more is that initial passive trials reduce cognitive load in the beginning of training. While our data supports this finding, we did not test another mixed condition where the passive trials take place at some other time in the study session, thus our data cannot speak to the benefits of *when* the passive trials happen or *how much* passive exposure is helpful; just that having a mix is better. For example, it is possible that having a mixture of both trial types helps learners encode different properties of the categories and if an active-passive group existed no differences would be found between that and passive-active. We have not addressed the extent to which the passive trials support performance on the active trials. How many passive exposures are optimal, and what is the relationship between the optimal number of exposures and the complexity of the learning domain? Additional research is needed to further understand and optimally utilize the passive-active synergy.

We have not perfectly controlled for the amount of motor response between the *active* and *passive* conditions. *Active* participants made two responses per trial, one to select an answer

and another to go to the next trial, whereas the *passive* participants only made one response per trial – to go to the next trial. These two could be equated by having the *passive* group clicking on the correct response. However, the level of engagement and motivation were not self-reported to be different between the *active* and *passive* condition in either experiments.

The *passive* groups in both experiments spent slightly longer time overall in the training modules, despite not actually having the information presented longer compared to the *active* groups. This likely reflected the general response to passive exposure training. One way to better equate the two conditions in time and trials is to give passive participants the exact same sequence and duration per trial as their active yoke. It is unlikely, however, that the pattern of results would change.

Furthermore, ECG interpretation was new learning for all participants, while recognition of Sine and Exponential transformations was a refresher for some. Although the literature on the expertise-reversal effect has shown a reversal effect occurs when level of prior knowledge interacts with the effectiveness of different instructional techniques (Kalyuga, 2007), where studying worked examples passively may be more effective than solving problems for novices in a domain, but the reverse is true for learners with a moderate amount of problem-solving knowledge in the domain (Kalyuga & Sweller, 2004), our result showed that the benefits of *passive-active* training generally held regardless of the level of novelty of the domains. This could be because most of our participants did not have high prior knowledge. More research is needed to understand whether the combination of passive and active classification produce similar learning gains and efficiency when learners already have high prior knowledge.

CONCLUSION

Passive presentations of category exemplars can act synergistically with active adaptive learning to elevate perceptual learning, transfer, and retention while decreasing training time. These effects were robust across two different complex real-world learning domains.

CHAPTER 5

Contrastive Comparison

INTRODUCTION

In her seminal book on perceptual learning, Eleanor Gibson (1969) placed particular emphasis on the idea that perceptual learning - as the process of differentiation, selection, and extraction of information - depends on the opportunity for stimulus comparison. In particular, she pointed out that "learning of differential properties should be facilitated by providing examples of contrasts along a dimension so as to define and assist isolation of the critical variable property" (Gibson, 1969, p. 99). Those differential properties are the distinctive features, the "relational contrasts" that distinguish one category of things from another. The opportunity to contrast instances that differ in structures should provide a good condition for differentiation learning, by allowing the learner to isolate the relevant features for distinguishing between things from the irrelevant ones (Gibson, 1969). Indeed, this theory has garnered much support in recent years.

When the features and relations within one stimulus are systematically matched to features and relations in the other stimulus ("aligned"), the similarities and differences across instances are made salient, allowing learners to pick out the distinguishing features and relations that both define a category and separate it from others (e.g., Markman & Gentner, 1997; Gentner & Gunn, 2001; Gentner, 2010). This comparison procedure can engage perceptual learning processes, leading to discovery of distinguishing features and or structural invariance, and to selective and fluent extraction of information in any particular domain (Gibson, 1969; Kellman, Massey & Son, 2009). Indeed, providing contrastive representations (i.e., comparing stimuli

belonging to different categories) for comparison has been found to help in discrimination learning for multiple domains involving perceptual learning such as radiology (e.g., Kok, Bruin, Robben, & van Merriënboer, 2012), cytopathology (Evered, Walker, Watt, & Perham, 2013), mathematics (e.g., Rittle-Johnson & Star, 2009), and ECG interpretation (Art, Brooks, & Eva, 2007), as well as in related learning contexts such as analogical reasoning (e.g., Holyoak, 2005; Gentner, 1983), category learning (Andrews, Livingston, & Kurtz, 2011; Ankowsky, Vlach, & Sandhofer, 2012), and schema acquisition (Gick & Paterson, 1992).

This raises an interesting possibility that by aligning the instances to-be-classified with another instance for contrast, the juxtaposition may make explicit the relevant information for discriminating between categories in PALMs. For example, one may guess that for ECG interpretation, the best contrasting example may be a tracing that shows no abnormalities (i.e., a normal ECG), so that when the normal features on both the normal tracing and the abnormal tracing are aligned to each other, the diagnostic features for abnormality can become the main difference between the two tracings. This saliency would influence visual attention and thus may make it easier for students to discriminate diagnostic-related information from irrelevant information. Alternatively, in graphical transformation learning, allowing students to contrast the canonical function with the transformed function may help them recognize the type of transformation that had occurred.

Is training with contrastive comparison *always* good? There is evidence suggesting that the benefit of contrastive comparison is dependent on task specifics such as feature variation and category structure (Ankowski, Vlach, & Sandhofer, 2012; Kok, Bruin, Robben, & van Merrienboer, 2012), which aspects of problem being compared (Rittle- Johnson & Star, 2009), learners' prior knowledge (Rittle-Johnson, Star, & Durkin, 2009), and experience of pretraining

(Braithwaite & Goldstone, 2014). For example, Kok et al. (2012) examined the effectiveness of training with and without contrastive examples on participants' ability to discriminate among 12 radiological appearances of lung and heart diseases on chest radiographs. They asked half of their medical student participants to compare diseased radiographs with images showing no abnormalities (normal images), while the other half only compared radiographs of patients with the same disease. Students who compared with normal images outperformed those who did not compare with normal images, but interestingly, this effect was found only for "focal diseases" that require attention to only one part of the image (such as a location of a specific mass or lesion) but not for diffuse diseases that require attention to be directed to various parts of the image (i.e., because the disease affects most of the chest). It is therefore possible that comparison has a stronger effect on discriminating information indicating focal diseases than on discriminating information indicating diffuse diseases. As Kok et al. (2012) remarked, "And when everything stands out, it does not stand out any more!" Contrastive experience can direct attention to the relevant information, making a focal disease standout, but for diffuse diseases, the whole image should become more salient. Consequently, attention is not directed to a specific location and discrimination of relevant information is not facilitated.

ECG interpretation is one typical domain in which there is a lot of information available and correct diagnoses require attention to multiple parts of the image rather than a specific location (Wood, Batt, Appelboam, Harris, & Wilson, 2013). Wood et al. (2013) has shown with visual search that expert cardiologists make a global or holistic assessment of the tracing, before locating the critical leads and quickly making the diagnosis. Even after they have already located the critical leads, they also cross-reference certain segments across leads. For example, they cross-referenced ST segments of inferior leads (left side of ECG) with the chest leads (right side

of the ECG), presumably to confirm that there were ST elevations compared with other leads and to look for reciprocal changes in the other leads. Given this "diffuse" nature of ECG interpretation, it is unclear whether comparison of contrastive examples would be effective.

Furthermore, in ECGs, there are variations in both the relevant and irrelevant leads. Ankowski, Vlach and Sandhofer (2012) have shown that the benefit of contrast is limited when categories have members that share few common features relevant for category membership and have many variations in features that are irrelevant for category membership. It is more effective when categories have members with many common features and variation in irrelevant features. The rationale was, when examples vary in multiple dimensions – including the diagnostic dimensions – they may provide too much overall variation to discover the relevant contrast. Detecting the relevant features for categorization may require aligning common features between simultaneously view representations (e.g., Gentner, 2005). This may be a difficult task for novices; when they view examples that differ in both the relevant target dimensions and in irrelevant dimensions, there may be too many similarities and differences to facilitate the type of alignment that leads to effective categorization.

Prior knowledge also plays a role in the effectiveness of contrastive comparisons. Contrastive comparisons have been shown to be effective at supporting learning for those with moderate prior knowledge (Rittle-Johnson & Star, 2007), but for those with low prior knowledge, the benefit is limited. Rittle-Johnson et al. (2009) studied students' procedural knowledge and flexibility in the equation solving task, and found the benefit of simultaneous comparison relative to sequential study of the same examples with middle-school students who varied in their prior knowledge of algebraic equation, but that students with low prior knowledge benefited more from sequential presentations. For those with moderate prior knowledge, one

possible reason for this advantage is the reduced memory constraints as compared to sequential presentation. This allows these learners to more effectively compare the two objects and extract the relevant information (Andrews, Livingston, & Kurtz, 2011). However, when learners have little prior knowledge, it is difficult to align unfamiliar examples because this unfamiliarity makes it hard for them to recognize which aspects to which they should attend (Gentner et al., 2003; Rittle-Johnson et al., 2009; Schwartz & Bransford, 1998).

Prior to comparing the similarities and differences between examples, the learner also needs to interpret each example to understand the importance of similarity and differences between examples. For novices, such a task can easily overload their working memory, as they must deal with many elements of information at once. In contrast, learners with more experience in a domain can use their existing knowledge structures to interpret and complete the task with better ease. Even though these studies by Rittle-Johnson, Star, and colleagues had students contrast alternative solution methods for the same problem and compare different problem types solved with the same solution method (isomorphic problems) with procedural solving knowledge as a goal of learning, similar principles may apply to perceptual classification learning.

How about when the category structure doesn't involve as much variations in the relevant and irrelevant features? In the math domain, Silva & Kellman (1999, unpublished data) manipulated the usage of contrastive instances during training and feedback for learning graphical transformations. They gave participants a fixed set of training trials (without learning to criterion). In each trial, participants matched graphs with equations that vary in their transformations from a canonical trigonometric function. They presented contrastive graphs, in which both the canonical function (e.g., $y = sin(x)$) and the function to-be-classified (e.g., $y = 2sin(x)$) were superimposed. In this way, they hoped to highlight the particular pattern

transformations relating the basic function to its variants.

When participants saw contrastive graphs during the learning trials *and* as feedback, they were better than the no-training control group on their ability to transfer their knowledge to correctly match complex combination functions like $y = cos(x) * log(x)$ or $y = sin(x) - exp(-x)$ to the appropriate graphs. However, contrastive graphs were not always helpful. Participants who studied only with contrastive graphs were not able to transfer their knowledge from one function family to another related one (i.e., those who studied Sine and Exponential functions were not able to classify Cosine and Logarithmic functions). The learners may become too dependent on the graphical exposure of the canonical function to be able to classify its variations.

Despite some intricacies, Silva & Kellman's findings suggested that while the comparison process can boost learners' extraction of relevant transformations in the learning set, allowing them to flexibly transfer what they learn to more complex situations, there might also be a negative effect on transfer when training only with contrastive instances.

Overview of Experiments 3 and 4

In Experiments 3 and 4, we began our examination of how we might capture the benefit of comparison for discrimination while minimizing participants' over-reliance on having a canonical instance available for contrast. We introduced a training condition in which participants received an equal mixture of *contrastive* and non-contrastive (*single*) learning instances.

There were some important differences in the way the contrastive experiences were presented in these two experiments. In Experiment 3, the *contrastive* presentation aligned a normal ECG half and an abnormal ECG half side-by-side. In Experiment 4, similar to Silva & Kellman (1999), the *contrastive* presentation presented a canonical function superimposed on the

same graph with its variant (e.g., $y = sin(x)$ and $y = 2sin(x)$ on the same graph). In Experiment 3, Normal was one of the categories to be learned, thus the *contrastive* normal ECGs varied throughout the training. In Experiment 4, the canonical functions were either $y = sin(x)$ or $y = exp(x)$ on each trial. These differences were inherent to the learning domain, and they allowed us to study the generalizability of the effect of *contrastive* presentations across domains.

In both domains, the *contrastive* presentation may prompt the learner to search for similarities and differences that explain how each specific instance diverges from the normal/canonical one. If this improves recognition of patterns in the transformations, it should be reflected in the transfer measures. However, the effectiveness of the *contrastive* experience may be dependent on the kind of discriminations to be learned and the learners' prior knowledge. The diagnostic features from ECGs are relatively more diffuse, whereas the graphical transformation the features are relatively more rule-based. Our ECG participants were novices to ECG interpretation, and most of our AlgGeo participants had some prior exposure but had forgotten much of the materials. In both experiments, we expected the *contrastive*-only experiences to be relatively less effective for learning than the *single* and *mixed* trainings, as the *contrastive*-only training might encourage participants to become over-reliant on the contrastive presentation, and/or present too much information to be helpful. The *mixed* condition, on the other hand, may alleviated some of these negative effects, and it may also bring about a change in practice that could produce a desirable difficulty for learners (e.g., Bjork, 1994, McDaniel & Butler, 2012) by forcing them to engage in deeper processing across training instances. This could make the learning process more difficult (and perhaps less efficient) but could thus lead to better learning outcomes. Thus, we hypothesized that training with *some* juxtaposition of learning instances can enhance the discovery of salient features to improve structure recognition in transfer situations.

Experiment 3

METHOD

Participants

Participants were 122 undergraduates from UCLA who have not had any prior ECG training (89 Female, mean age = 20.69). We analyzed and reported results from 90 of those participants who reached learning criteria (64 Female, mean age = 20.45).

Design

Participants were randomly assigned into one of three training conditions: *single*, *contrastive*, and *mixed*.

Materials

We used only ECG halves in this experiment rather than the full 12-lead ECGs. From the original 250 ECG traces, there were 250 half-ECGs from the left side of the tracings, and 250 half-ECGs from the right side of the tracings. This was possible because the heart patterns we chose did not require participants to view the entire 12-lead ECG to make a classification. Each ECG contained sufficient diagnostic information in either the left half or the right half (depending on the particular diagnosis; e.g., the right half for LAD and RAD, left half for Anterior STEMI).

We distinguished between two types of training trials: *single* and *contrastive*. The *single* trial contained a single half-ECG and the same 7 answer options as those in Experiment 1. The *contrastive* trial contained 2 half-ECGs (from the same side): a single half-ECG to-be-classified and a Normal half-ECG for contrast, and 7 answer options. To indicate which half was which, in both trial types, if the to-be-classified ECG was a left half of an ECG (Left-ECG), it always appeared on the left side of the screen, and if it was a right half of an ECG (Right-ECG), it

always appeared on the right side of the screen. The more informative half for each pattern was always used as the to-be-classified ECGs. There was also a pink stripe marking on the left of each Left-ECG and on the right of each Right-ECG to indicate which side it originally came from. On *contrastive* trials, the Normal half-ECG always came from the same side of the to-be-classified half-ECG, so that there were either two Left-ECGs or two Right-ECGs on each *contrastive* trial. The Normal half-ECGs were randomly selected from the Normal category on each *contrastive* trial. *Figure 3.1* shows a sample *single* trial and its feedback, and *Figure 3.2* shows a sample *contrastive* trial and its feedback. All PALMs used the same pool of ECGs.

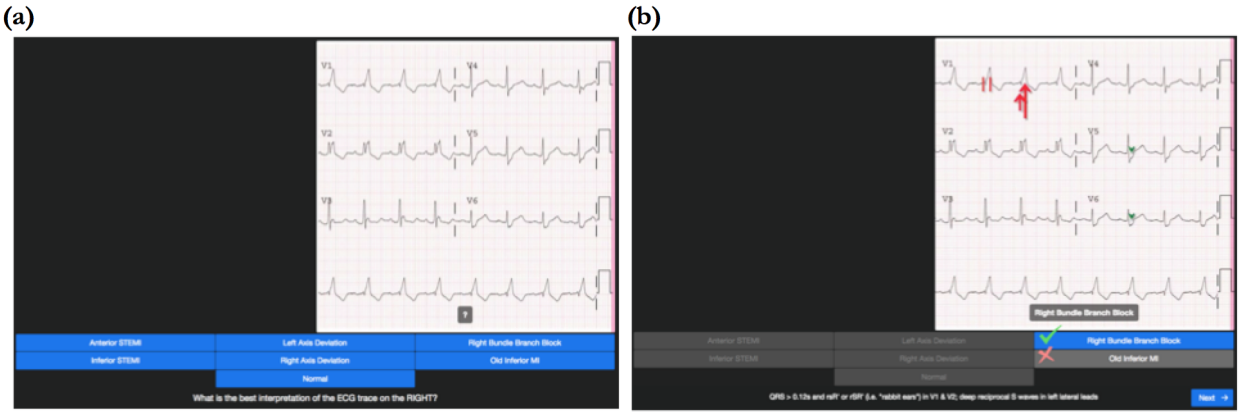


Figure 3.1. (a) Sample *single* trial and (b) its feedback

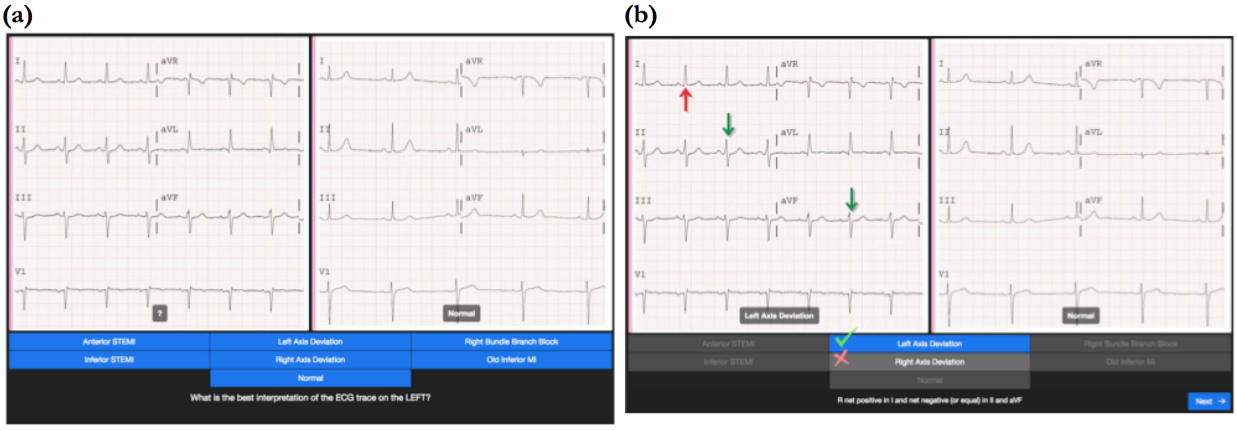


Figure 3.2. (a) Sample *contrastive* trial and (b) its feedback

Procedure

The procedure was the same as that used in Experiment 1, except the three different training conditions.

Training

In the *single* condition, participants received only single trials. In the *contrastive* condition, participants received only contrastive trials. In the *mixed* condition, participants received an equal mixture of single and contrastive trials. These trials were randomly selected on each trial. All three conditions shared the learning criteria as used in Experiment 1 (4/4 consecutive correct answers for each heart pattern, each under 15 seconds).

Assessments

Assessments were the same as those used in Experiment 1.

Overview of Analyses and Expected Results

The three groups did not differ on quiz performance or any survey measures. They differed slightly in module completion rate (79% of the *single* group, 75% of the *contrastive* group, and 68% of the *mixed* group completed within the allotted time). Here we report results from the first 30 participants in each condition who have reached learning criterion. The same general patterns of results were found when we included *all* participants who have attempted the modules. *Appendix F.1* contains results from all participants. All assumptions for ANCOVA were met for all dependent variables: (1) independence of the covariate and the treatment effect, F's < 1, p's > .05; and (2) homogeneity of regression slopes (the correlation between pretest and the posttest gains were roughly equal across conditions), F's < 1, p's < .05.

Similar to Experiment 1, because participants were trained toward learning criteria, we expected substantial learning gain and retention from all three training conditions. We expected

the *mixed* and *single* conditions to outperform the *contrastive* condition on transfer, and as for the difference between *mixed* and *single* conditions, if there are advantages to having 50% *contrastive* trials, then the *mixed* condition should enhance learning and retention better than the *single* condition. Along the same line, in terms of efficiencies, if *contrastive* displays make it easier for learners to recognize relevancies, those in the *contrastive* and *mixed* conditions should complete the module faster than those in the *single* condition. However, the *mixed* presentations could introduce a desirable difficulty, so that learners might require more time or effort in the *mixed* condition, but may acquire the biggest gain in transfer scores.

An alternative possibility was that the *contrastive* displays pose a cognitive load issue for novices, which may eliminate the benefit of contrastive comparisons. If this were the case, we would expect the *single* condition to outperform the *contrastive* and *mixed* conditions on all measures.

RESULTS

Efficiency

Figure 3.3a shows the efficiencies as computed by trials and *Figure 3.3b* as computed by time. The *single* and *mixed* conditions outperformed the *contrastive* condition on both measures of efficiency. This observation was confirmed by the analyses.

Efficiency by Trials

A 2 phase (pre-post, pre-delayed) x 3 conditions (*single*, *contrastive*, *mixed*) ANCOVA with pretest accuracy as the covariate showed a marginally significant phase x condition interaction, $F(2,86) = 2.39$, $p = .097$, $\eta^2_p = .05$. There were no condition differences on pre-post efficiency, but on pre-delayed efficiency, the *single* ($M = .0023$, $SD = .0021$) and *mixed* ($M =$

.0017, *SD* = .0014) conditions produced greater and marginally greater efficiency than the *contrastive* condition with medium effect sizes (*M* = .0009, *SD* = .0018), t(58) = 2.84, *p* = .006, *d* = .73, and *t*(58) = 1.97, *p* = .05, *d* = .51, respectively. The *single* and *mixed* conditions did not differ on pre-post trial efficiency, *t*(58) = 1.52, *p* = .13, or on pre-delayed trial efficiency, *t*(58) = 1.33, *p* = .19. There were no other condition differences, and no main effect of phase, *p*'s > .10.

There was a main effect of the covariate on time efficiencies, *F*(1,86) = 8.71, *p* = .004, η^2_p = .09. Pretest accuracy did not correlate with the pre-post trial efficiency, *r*(90) = -.16, *p* = .15, but it did correlate strongly with the pre-delayed trial efficiency, *r*(90) = -.49, *p* < .001. The same pattern of condition differences was found when analyzed without correcting for pretest variations.

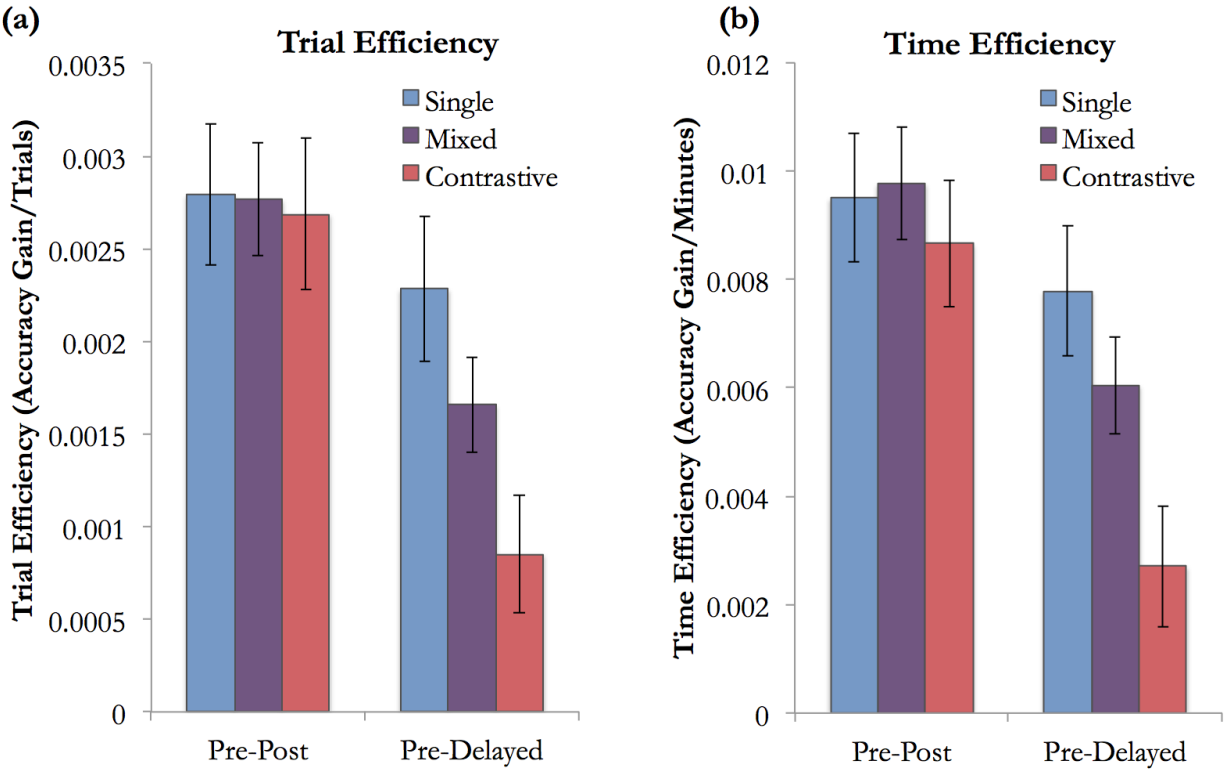


Figure 3.3. Training efficiencies (a) by trial and (b) by time. Error bars are ± 1 standard error.

Efficiency by Time

Efficiency by time showed identical patterns. While there was no differences among condition on pre-post efficiency, both the *single* (M = .008, SD = .007) and *mixed* (M = .006, SD = .005) conditions produced greater pre-delayed efficiency than the *contrastive* condition (M = .003, SD = .006) with medium effect sizes, $t(58) = 3.11$, $p = .003$, $d = .80$, and $t(58) = 2.34$, $p = .02$, $d = .60$, respectively. The *single* and *mixed* conditions did not differ on pre-delayed time efficiency, $t(58) = 1.16$, $p = .25$.

Accuracy

Figure 3.4a shows accuracy by condition. All three conditions produced strong learning gains and transfer to novel instances at immediate posttest, and long-term retention seen at a one-week delay, $F(2,174) = 141.13$, $p < .001$, $\eta^2_p = .62$. Across all conditions, immediate posttest accuracy (M = .60, SD = .15) and delayed test accuracy (M = .47, SD = .15) were significantly greater than pretest with very large and large effect sizes (M = .29, SD = .14), $t(89) = 16.25$, $p < .001$, $d = 2.17$, and $t(58) = 8.92$, $p < .001$, $d = 1.23$, respectively. The drop from immediate posttest to delayed test was also statistically significant, $t(89) = 7.22$, $p < .001$, $d = .87$.

Accuracy results shared the same pattern of condition differences as efficiency. The *single* and the *mixed* conditions produced overall better retention after the delay than the *contrastive* condition, $F(4,174) = 3.75$, $p = .006$, $\eta^2_p = .08$. At pretest and immediate posttest, there were no differences across conditions (mean ranged .27 – 32 at pretest, and .58 - .62 at immediate posttest), $t(58) < 1.1$, p's > .20. However, at delayed test, the *single* condition (M = .52, SD = .13) outperformed the *contrastive* condition with a medium effect size (M = .41, SD = .18), $t(58) = 2.79$, $p = .007$, $d = .70$. The *mixed* condition (M = .49, SD = .13) also performed marginally higher than the *contrastive* condition at delayed test, also with a medium effect size,

$t(58) = 1.97$, $p = .05$, $d = .51$. *Single* and *mixed* did not differ at delayed test, $t(58) < 1$, $p > .20$. There was no other effect, p's > .10.

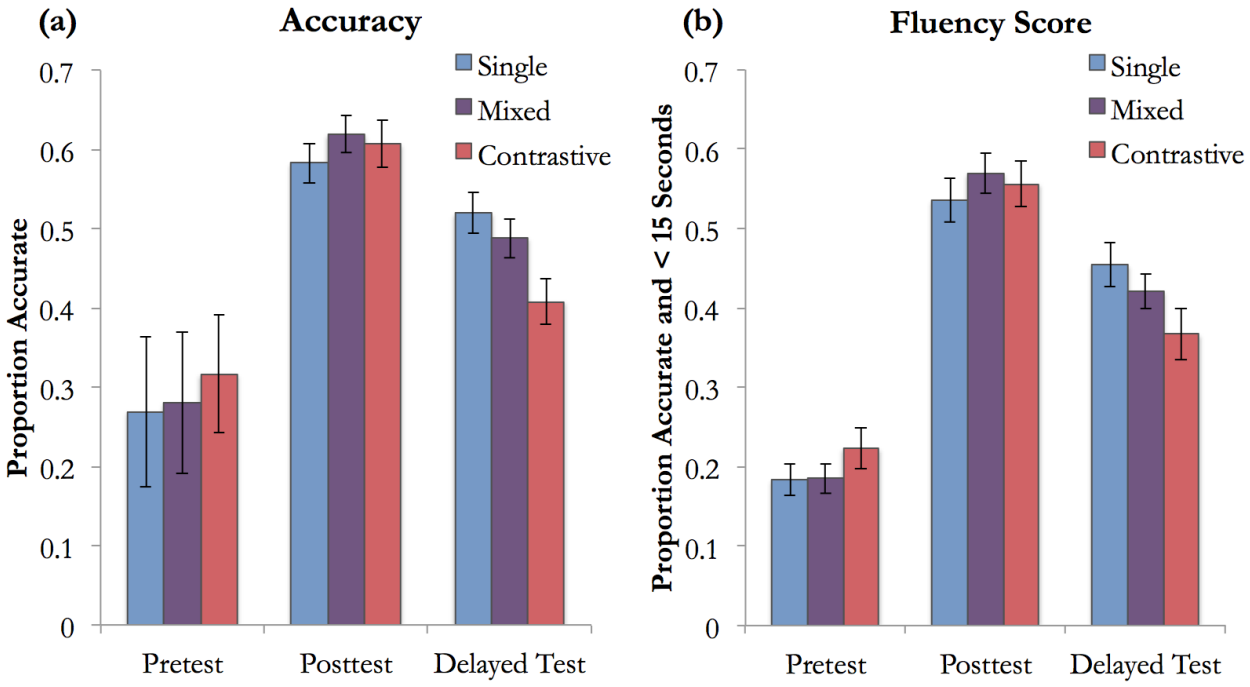


Figure 3.4. (a) accuracy and (b) fluent accuracy in Experiment 3.

Error bars are ± 1 standard error.

Accuracy Gain

Accuracy gain showed the same pattern. There were no condition differences in pre-post gain, $t(58) < 1.1$, $p > .20$, but both the *single* ($M = .25$, $SD = .18$) and *mixed* ($M = .21$, $SD = .16$) conditions produced significantly greater pre-delayed gain than the *contrastive* condition with medium to large effect sizes ($M = .09$, $SD = .21$), $t(58) = 3.13$, $p = .003$, $d = .82$, and $t(58) = 2.39$, $p = .02$, $d = .64$, respectively. *Single* and *mixed* did not differ on pre-delayed test gain, $t(58) < 1$, $p > .20$.

Fluency

Figure 3.4b shows the mean fluent accuracy by condition. The pattern was similar to that of accuracy, with the exception that the *mixed* condition was not reliably better than the

contrastive condition at delayed test, $t(58) = 1.40$, $p = .17$. The *single* condition had higher fluent accuracy than the *contrastive* condition at delayed test, $t(58) = 2.08$, $p = .04$, $d = .51$. In terms of pre-delayed fluency gain, both of the *single* and *mixed* conditions were reliably better than *contrastive* with medium effect sizes, $t(58) = 2.72$, $p = .009$, $d = .70$, and $t(58) = 2.20$, $p = .032$, $d = .62$, respectively. The *single* and *mixed* conditions did not differ on any fluency measures, $t(58) < 1$, $p > .20$. There were no condition differences in RTc, p's > .10. *Appendix F.2* contains more details of these analyses.

Progression of Learning

Table 3.1 displays the training means by conditions. There were no reliable differences between conditions on the number of trials and amount of time spent on the training module, nor on accuracy and fluent accuracy (p's > .10). Thus, differences on efficiency were driven by differences in accuracy gain.

Conditions	Trials Completed	Minutes on Training	Training Accuracy	Training Fluent Accuracy	Training RTc
Single	136.0 (10.58)	36.8 (2.42)	.57 (.02)	.52 (.02)	7.85 (.37)
Contrastive	139.9 (10.42)	36.8 (13.61)	.56 (.02)	.49 (.02)	8.95 (.39)
Mixed	141.0 (9.06)	39.6 (2.61)	.54 (.02)	.49 (.02)	8.40 (.34)
- Single trials	70.3 (4.69)		.55 (.02)	.51 (.02)	7.57 (.31)
- Contrastive trials	70.7 (4.66)		.53 (.02)	.46 (.01)	9.28 (.40)

Table 3.1. Training means by condition (standard errors in parentheses). Performance on single and contrastive trials in the *mixed* condition are shown separately.

Interestingly, however, there were significant differences in RTc between the *contrastive* and *single* training trials (*Figure 3.5*). A 4 quartile x 3 condition on RTc confirmed a marginally significant main effect of condition, $F(2,87) = 2.74$, $p = .07$, $\eta^2_p = .06$. The *contrastive* condition spent about a second longer per correct trial on average ($M = 8.87$ seconds, $SD = 2.12$) than the *single* condition ($M = 7.72$ seconds, $SD = 2.00$), $t(58) = 2.15$, $p = .04$, $d = .56$. This was consistent across the first three quartiles. At the 1st quartile, *contrastive* condition took 11.50 seconds on average ($SD = 3.04$) for each correct answer compared to the *single* condition taking 9.6 seconds ($SD = 2.92$), $t(58) = 2.43$, $p = .02$, $d = .64$. This was apparent at the first two blocks of the training with medium effect sizes, $t(56) = 2.35$, $p = .02$, $d = .62$, and $t(58) = 2.07$, $p = .04$, $d = .54$[8]. This RTc difference between the *contrastive* and *single* condition were marginally significant throughout the 2nd and 3rd quartiles with a medium effect size, $t(58) = 1.94$, $p = .06$, $d = .50$, and $t(58) = 2.05$, $p = .05$, $d = .53$, respectively, and did not persist in the last quartile of the training, $t(58) = 1.23$, $p > .10$.

The *mixed* group provided a within-subject comparison, and indeed, they also spent longer to get each question correctly on *contrastive* trials ($M = 9.28$ seconds, $SD = 2.21$) than on *single* trials on average ($M = 7.57$ seconds, $SD = 1.71$), $t(29) = 7.49$, $p < .001$, $d = .87$. The difference in processing time was small and marginally significant in the 1st quartile (11.16 seconds on contrastive trials vs. 9.91 seconds on *single* trials), $t(29) = 1.96$, $p = .06$, $d = .40$, but was larger and persisted throughout the later 3 quartiles with medium to large effect sizes, $t(29) > 4.81$, $p < .001$, $d = .66$ to $.92$.

The *mixed* group took similar amount of time on correct contrastive trials as the *contrastive* group and similar amount of time on correct single trials as the *single* group, p's >

[8] Note the difference in degrees of freedom. This was because at block 1, there were participants who did not any questions correctly.

.10. The *contrastive* condition did not differ from the *mixed* condition and the *mixed* condition did not differ from the *single* condition on overall RTc, *p*'s > .10.

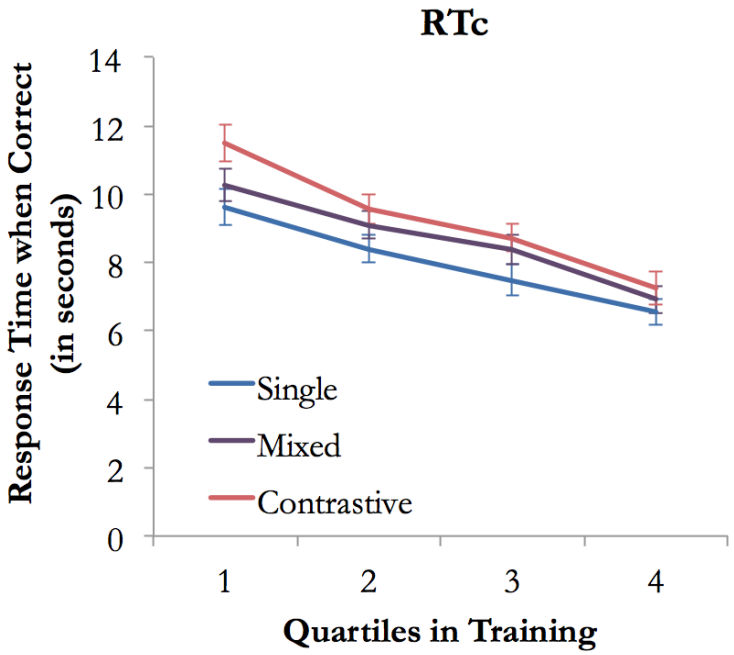


Figure 3.5. RTc by training quartiles. Error bars are ± 1 standard error.

Survey Data

Interestingly, participants seemed to favor the *contrastive* experiences. One-way ANOVAs[9] confirmed a main effect of condition on enjoyability ratings, $F(2,86) = 3.78$, $p = .03$. The *mixed* ($M = 4.38$, $SD = 1.18$) and *contrastive* ($M = 4.37$, $SD = 1.22$) conditions were considered to be significantly more enjoyable than the *single* condition ($M = 3.57$, $SD = 1.50$) with medium effect sizes, $t(57) = 2.31$, $p = .02$, $d = .59$, and $t(58) = 2.27$, $p = .03$, $d = .60$. There were no differences between the *contrastive* and *mixed* conditions, $p > .10$.

Participants in the *mixed* condition ($M = 4.45$, $SD = 1.18$) and the *contrastive* condition rated the primer to be more helpful than those in the *single* condition ($M = 3.60$, $SD = 1.40$),

[9] Survey data from one participant in the *mixed* condition was not recorded properly, so the following analyses were conducted with 89 participants.

$t(57) = 2.51$, $p < .02$, $d = .66$, and $t(57) = 1.89$, $p = .06$ (marginal), $d = .48$, respectively. The *mixed* group ($M = 5.10$, $SD = .92$) also rated their module to be marginally more helpful than the *single* group ($M = 4.70$, $SD = 1.51$), $t(57) = 1.94$, $p = .06$, $d = .50$, and themselves to be marginally more engaged and motivated in the training ($M = 4.90$, $SD = 1.15$), than the *single* group ($M = 4.30$, $SD = 1.54$), $t(57) = 1.69$, $p = .096$, $d = .44$. These differences had small to medium effect sizes. The *contrastive* group did not differ from the other two on these survey responses, p's > .10.

On a metacognitive survey question, we described the condition manipulation and asked participants to predict the posttest performance among conditions. Interestingly, most participants (43.3%, or 39 of the 90 participants) and more than half of the *single* group (16/30) chose the *contrastive* training. 22.2% (20 out of 90) chose the *mixed* condition, and another 22.2% chose the *single* condition, and the remaining 12.2% (11 out of 90) thought all three conditions would be equally effective. The three groups did not differ on any other self-rating measures, p's > .10.

DISCUSSION

The three conditions produced equal learning gains and efficiencies at immediate posttest, but the *single* and *mixed* conditions produced better performance than the *contrastive* condition at delayed test.

Prior research suggested that contrasting examples with the normal/canonical patterns should help learners discriminate diagnostic features from irrelevant features (Gentner & Gunn, 2001), resulting in better mental representation of the diagnostic category. Notably, however, training with only contrastive examples proved to be a disadvantage in this study. Although the *contrastive* condition did equally well as the *single* and *mixed* conditions at immediate posttest,

its disadvantage was prominent at delayed test, when the *contrastive* group experienced more forgetting.

Interestingly, during the training, participants in the *contrastive* condition needed more time to reach the correct answers. Those in the *mixed* condition also took longer on *contrastive* trials than *single* trials. A positive interpretation of this result might be that contrasts facilitated additional processing and enhanced learning. However, the learning data do not support this interpretation. A less positive account is that *contrastive* trials may have provided too much information on the screen to be useful. Participants may not have tried to compare and contrast as much as we had expected them to. Indeed, participants have spontaneously reported that they did not paying attention to the Normal ECG after some time into the module. Because the ECG halves appeared on different sides of the screen depending on the relevant features they contained, the extra second needed on *contrastive* trials may simply to figure out which ECG was to be classified and which was the Normal ECG, not to scan across the images to find similarities and differences. This may have been the case for most participants, as supported by similar response times on correct answers in the last training quartile compared to the *single* condition. This suggested that while the *contrastive* trials provided more information to be processed, it was not necessarily engaging learners in a way that would be helpful for retention. The *mixed* condition did not show a learning advantage over the *single* condition.

ECG interpretation is a typical domain in which novices tend to attend to information based on conspicuity than on relevance, even if this conspicuous information is not relevant (Wood et al., 2013). The disadvantage of the *contrastive* condition is consistent with the finding from Kok et al. (2013), which suggested that for domains like ECG in which the relevant information is "diffuse" throughout the image and there are likely too many similarities and

differences that may be difficult for novices to parse. In these instances, comparisons of contrasts are unlikely to be helpful (Kok et al., 2013). Furthermore, our participants had low prior knowledge of ECG - even after the primer, they averaged less than 30% at pretest. With little prior knowledge, it could have been difficult for them to know what to look for, where and how to look for them (Gentner, Loewenstein, & Thompson, 2003; Rittle-Johnson, Star, & Durkin, 2009; Schwartz & Bransford, 1998). Thus, comparison may be too overwhelming to significantly improve learning. This finding is also consistent with findings from the classic research on aptitude–treatment interactions (Snow, 1992) and more recent research on expertise reversal effects (see Kalyuga, 2007, for a review), which suggested that students with low prior knowledge benefited more from sequential presentations (that were similar to that seen in the *single* condition, Rittle-Johnson et al., 2009).

The *single* and *mixed* conditions did not differ reliably on any measures, except that participants in the *mixed* condition rated their training experiences to be more enjoyable, engaging, and the module to be more helpful than the *single* condition. Mixing the two trial types may not have led to better accuracy or efficiency at a delay than the *single* condition, but the combination of trial types (or perhaps the presence of the *contrastive* trials) produced enough variation that allowed the training to be perceived as more enjoyable, engaging, and helpful. In general, participants preferred having the *contrastive* training experience and thought that it was marginally more enjoyable for learning than the *single* training, yet, our findings showed that having only *contrastive* trials comparisons was damaging for learning. This is consistent with research on metacognitive illusions of learning (e.g., Bjork 1994; Castel, McCabe, & Roediger, 2007).

Experiment 4

In this experiment, we asked, do the same disadvantages of contrastive comparisons apply to graphical transformation learning?

METHOD

Participants

72 Amazon Mechanical Turk workers (38 Female, mean age = 31.81, *SD* = 9.42) completed all parts of the study for monetary compensation. The eligibility requirements were identical to that used in Experiment 2. Please see *Appendix G* for more demographic information.

Design

Participants were randomly assigned into one of three training modules based on their pretest accuracy (similar to Experiment 2). The three training modules were identical in every way except for the type of graphs shown on each trial: (1) only *single* graphs, (2) only *contrastive* graphs, and (2) a random equal mixture of the two (*mixed*).

Materials

Training

Graphs were presented either as *simple* or as *contrastive* displays. The *simple* version of the graph consisted of just to-be-classified function, such as $y = sin(2x)$, displayed as a thick blue line. Each contrastive graph contained both the to-be-classified function (i.e., $y = sin(2x)$), displayed as a thick blue line, and the canonical function (i.e., $y = sin(x)$), displayed as a superimposed dotted gray line. *Figures 4.1a* and *b* show a sample *contrastive* trial and its feedback. *Contrastive* graphs were used as feedback on *single* trials.

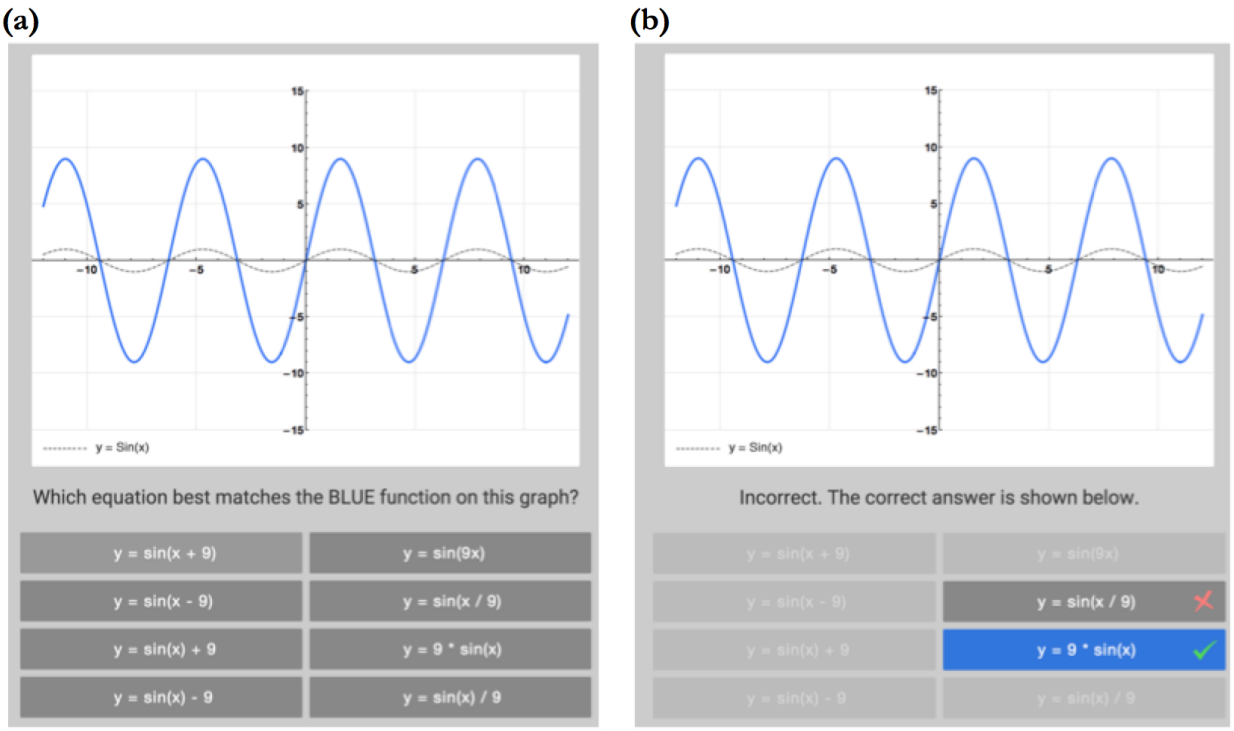


Figure 4.1. (a) Sample *contrastive* trial and (b) its feedback.

There were two other changes to the materials used in Experiment 2: (1) We added more graphs to equate the variations among categories so that there were 9 examples per transformation subtype, creating a total of 144 single graphs and 144 contrastive graphs; (2) Each learning trial in these modules contained a graph and 8 answer choices representing 8 transformation subtypes for each function family (rather than just 4 answer choices). This was to better estimate participants' ability to discriminate between the different transformation subtypes.

Assessments

The assessment questions were the same sets of 28 questions as those used in Experiment 2. However, unlike Experiment 2, the trained items (TI), trained fuctions, novel items (TF/NI),

and untrained function (UF) items presented 8 answer choices. The combination function (CF) items remained the same as used in Experiment 2 with 4 answer choices each.

Procedure

The procedure was identical to those used in Experiment 2.

In the *single* module, each trial presented a single graph and 8 answer choices. After each response, the graph is replaced with its contrastive version. In the *contrastive* module, each trial presented a contrastive graph, including those used as feedback. In the *mixed* module, trials containing *simple* and *contrastive* graphs were randomly interleaved, and the feedback always showed the *contrastive* version. All participants learned toward the same learning criteria (3/3 correct per transformation type, each in under 15 seconds).

Overview of Analysis and Expected Results

Similar with Experiment 2, we only collected and analyzed data from participants who have completed all phases of the study (N = 24 per condition), all of whom did not experience technical difficulties and did not self-report to have looked up the materials at any point during the study. There were 16 others in the *single* condition, 18 in the *contrastive* condition, and 24 in the *mixed* condition who started their PALM but dropped out during the module.

We expected the same results as those in Experiment 3. Specifically, we expected the *single* and *mixed* conditions to perform better than the *contrastive* condition. We also expected to replicate findings from Silva & Kellman (1999) that *contrastive* practice may reduce transfer performance on certain types of transfer classifications. The *mixed* condition, on the other hand, may provide the benefits of contrastive experience while preventing learners from becoming over-reliant on always having the canonical function available for contrast (as seen earlier with Silva & Kellman, unpublished data). Therefore, the *mixed* condition could produce robust gains

on both near and remote transfer test items when compared to participants in the *contrastive* condition.

RESULTS

Efficiency

Efficiency by Trial

Figure 4.2a displays the overall trial efficiency by condition. The *contrastive* condition had numerically lower trial efficiencies than the *single* and *mixed* conditions, but the differences were not statistically significant, $t(58) = 1.49$, $p = .14$ and $t(58) = 1.45$, $p = .15$, respectively. There were no other reliable condition differences on the overall trial efficiency, $t(58) < 1$, p's > .20.

By Assessment Item Types

When calculated with accuracy gain on trained items (TI), the only notable condition effect was that the *single* condition's pre-delayed TI efficiency was marginally higher than that of the *contrastive* condition ($M = .0013$, $SD = .0016$ vs. $M = .0006$, $SD = .0011$, respectively), $t(46) = 1.66$, $p = .10$, $d = .48$. Also, when calculated with accuracy gain on untrained functions (UF), the *single* condition was marginally more efficient than the *contrastive* condition ($M = .0010$, $SD = .0014$ vs. $M = .0002$, $SD = .0013$, $t(46) = 2.00$, $p = .05$, $d = .58$). There were no other condition differences on other item type, $t(58) < 1.3$, p's > .20.

Efficiency by Time

Figure 4.2b displays the time efficiency by condition. Time efficiency showed a slightly different pattern, but there were no reliable condition differences, except when considered with novel items of trained functions (TF/NI) accuracy. In terms of TF/NI time efficiency, the *mixed* condition ($M = .0015$, $SD = .0014$) did better than the *single* condition ($M = .0012$, $SD = .0015$)

with a small effect size, $t(46) = 2.06$, $p < .05$, $d = .47$. There were no other condition differences, p's > .10. *Appendix H* contains more detailed analyses.

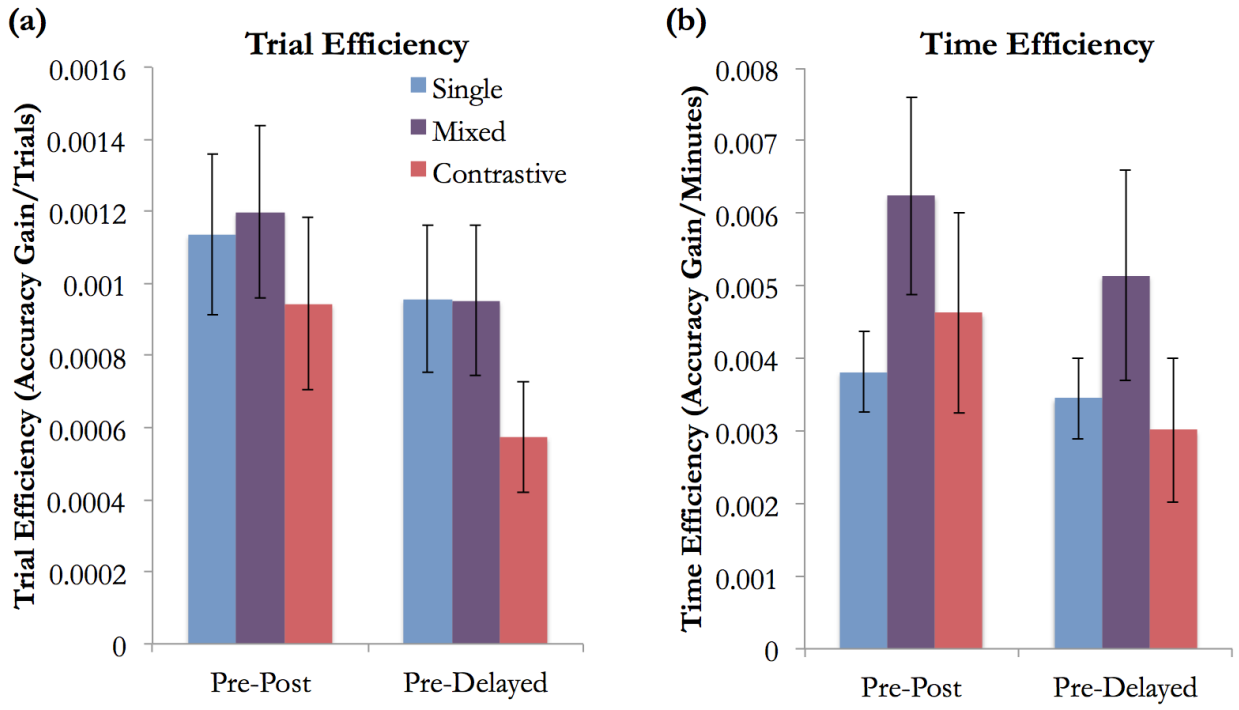


Figure 4.2. Efficiency (a) by trial and (b) by time for all items. Error bars are ± 1 standard error.

Accuracy

All Items

Figure 4.3a shows the mean overall accuracy on the assessments by condition. There was a main effect of phase, $F(2,138) = 78.04$, $p < .001$, $\eta^2_p = .53$. Across conditions, participants showed robust improvements on all items from pretest to immediate posttest (21% to 39%), $t(71) = 10.52$, $p < .001$, $d = 1.51$, and from pretest to delayed test (21% to 36%), $t(71) = 9.12$, $p < .001$, $d = 1.21$. There was a small 3% drop between immediate posttest and delayed test, $t(71) = 2.38$, $p < .05$, $d = .22$. The gain between pretest and delayed test was reliable for all item types (see *Appendix H* for more details of these analyses).

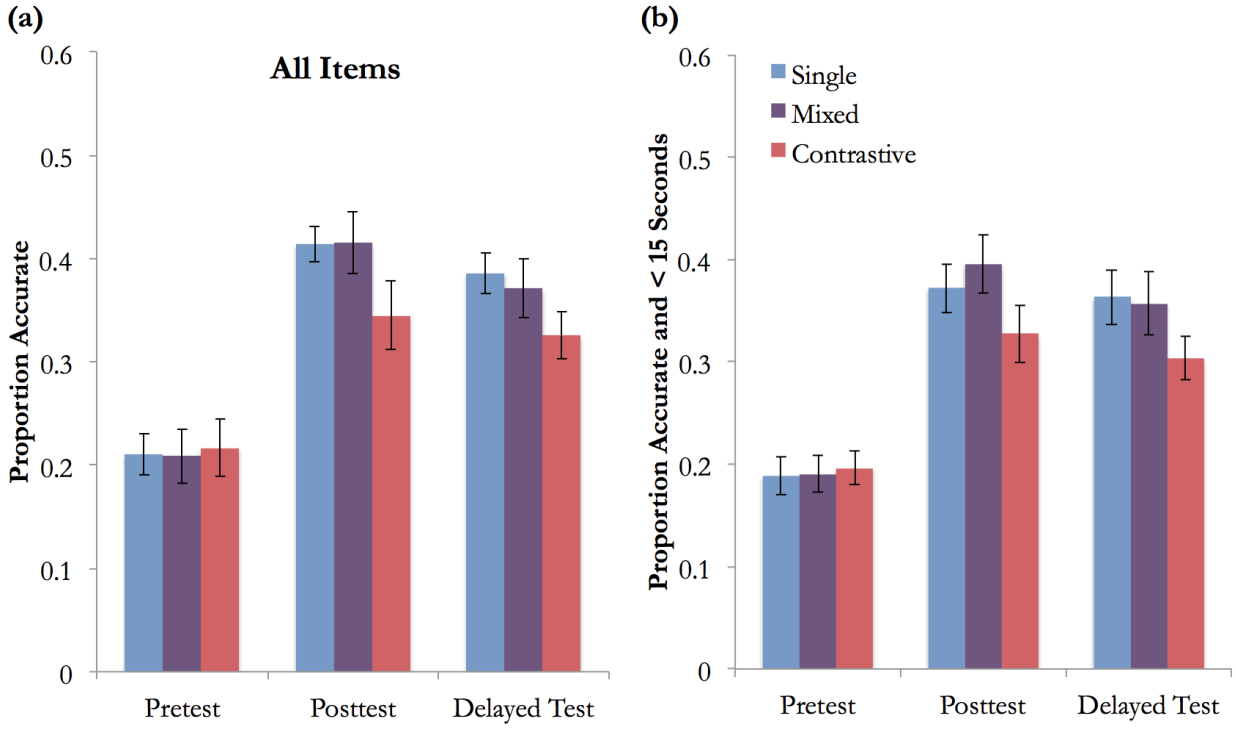


Figure 4.3. Mean (a) accuracy and (b) fluent accuracy on all assessment items. Error bars are ± 1 standard error.

There were no main effect of condition and no phase x condition interaction, *p*'s > .10. Our planned comparisons, however, confirmed that at immediate posttest, the *single* (M = .41, SD = .13) and *mixed* groups (M = .42, SD = .15) scored marginally higher than the *contrastive* group (M = .35, SD = .14) with small-medium effect sizes, $t(46) = 1.77$, $p = .08$, $d = .51$, and $t(46) = 1.71$, $p = .10$, $d = .49$, respectively. At delayed test, *single* also had marginally higher accuracy (M = .39, SD = .13) than *contrastive* (M = .33, SD = .11), $t(46) = 1.67$, $p = .10$, $d = .48$, but the effect was small. There were no other differences across conditions, *p*'s > .10. Accuracy gain confirmed the same patterns.

By Assessment Item Types

Figure 4.4a shows the average trained items (TI) accuracy by condition. On TI, there was a marginal main effect of condition, $F(2,69) = 2.81$, $p = .07$, $\eta^2_p = .08$, with the *single* condition doing better overall than the *contrastive* condition, $t(46) = 2.36$, $p = .02$, $d = .68$. This was driven by condition differences on Sine TI, $t(46) = 2.38$, $p < .05$, $d = .68$, less so on Exponential TI items, $p > .10$. There were no other condition differences, p's $> .10$. However, all three conditions showed robust improvements from pretest to immediate posttest, and from pretest to delayed test on TI, ranging from medium to very large effect sizes, $t(23) > 3.31$, p's $< .001$, $d =$ .68 to 1.72. A summary of these analyses is shown in *Table H.1* in *Appendix H.*

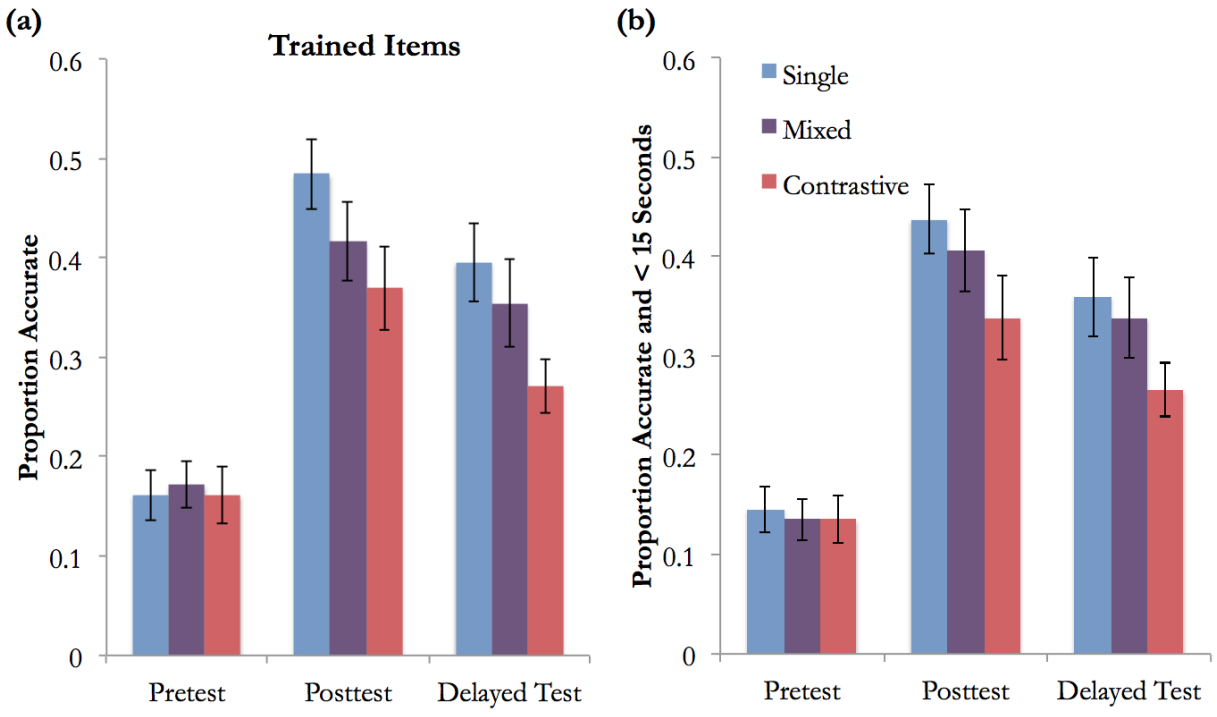


Figure 4.4. Mean (a) accuracy and (b) fluent accuracy on Trained Items.

Error bars are ± 1 standard error.

Figure 4.5a shows the mean accuracy on novel items of trained functions (TF/NI) by condition. At immediate posttest, the *mixed* condition ($M = .48$, $SD = .19$) did marginally better

than the *contrastive* condition with a small effect size (M = .38, SD = .19), $t(46) = 1.91$, $p = .06$, $d = .47$, but not at delayed test, $p > .10$. All three conditions showed strong learning gains from pretest to immediate posttest, and from pretest to delayed test, with medium to very large effect sizes, $t(23) > 3.15$, $p < .01$, $d = .64$ to 1.42. This was driven by condition differences on Exponential TF/NI; at immediate posttest, both the *mixed* and *single* conditions outperformed the *contrastive* condition with medium effect sizes, $t(46) = 2.29$, $p = .03$, $d = .66$, and $t(46) = 2.62$, $p = .01$, $d = .76$, respectively. However, there were no condition differences at delayed test, and no other condition differences, $t(46) < 1$, p's > .20. On Sine TF/NI, the *mixed* condition had numerically higher delayed test means than both the *contrastive* and *single* at immediate posttest and delayed test, but the differences were not reliable, $t(46) < 1$, $p > .20$ at immediate posttest, and $t(46) < 1.6$, p's = .12 at delayed test. *Figure 4.6* shows the mean accuracy on Exponential TF/NI and on Sine TF/NI items.

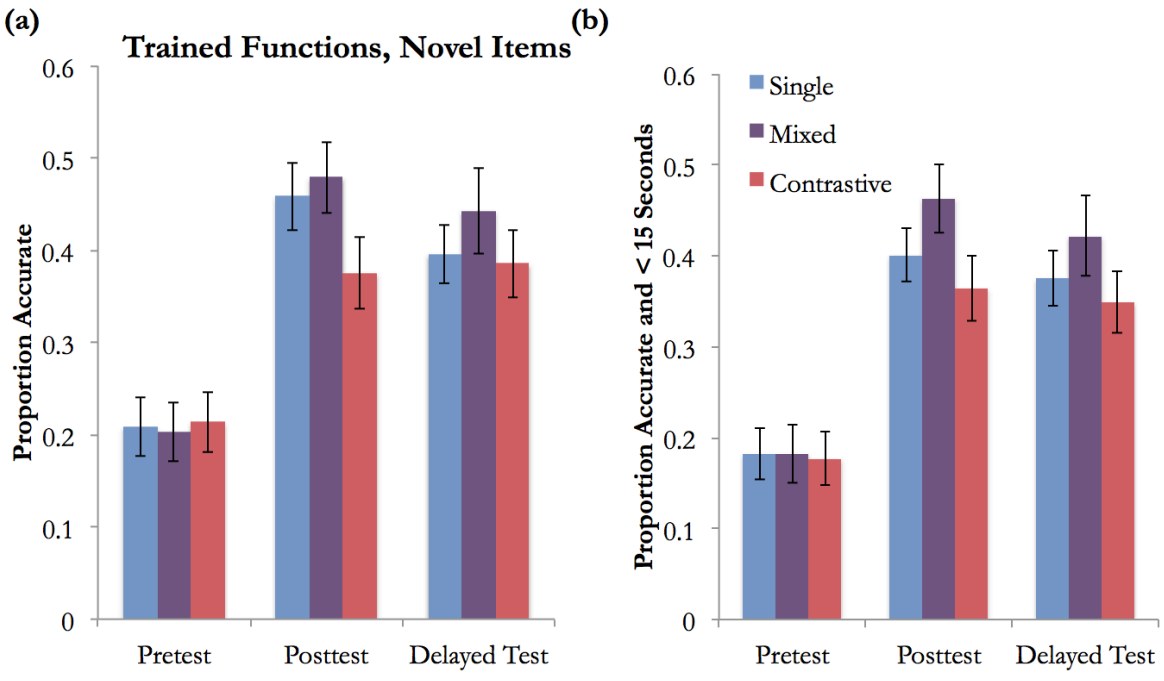


Figure 4.5. Mean (a) accuracy and (b) fluent accuracy on Trained Functions, Novel Items. Error bars are ± 1 standard error.

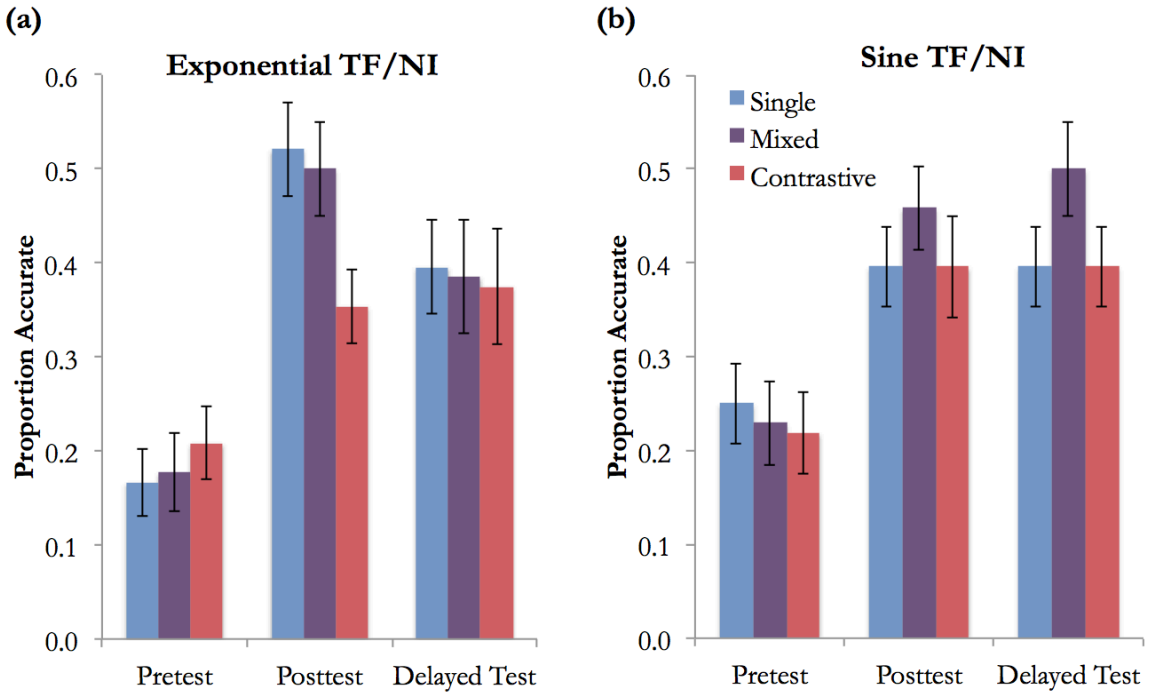


Figure 4.6. Mean accuracy on (a) Exponential TF/NI and (b) Sine TF/NI items. Error bars are ± 1 standard error.

Figure 4.7a shows the average untrained functions (UF) accuracy by condition. On UF items, the *single* condition exhibited marginally higher learning gain than the *contrastive* condition with medium effect size, $t(46) = 1.98$, $p = .05$, $d = .52$. Interestingly, the *single* and *mixed* conditions showed reliable pre-post and pre-delayed improvements on UF accuracy with medium to large effect sizes, but the *contrastive* condition did not (pre-post: *single*: $t(23) = 3.84$, $p < .01$, $d = .99$; *mixed*: $t(23) = 3.18$, $p < .01$, $d = .95$; *contrastive*: $t(23) = 1.62$, $p = .12$; Pre-delayed: *single*: $t(23) = 4.74$, $p < .001$, $d = 1.09$; *mixed*: $t(23) = 1.90$, $p = .07$, $d = .55$; *contrastive*: $t(23) = 1.59$, $p = .13$).

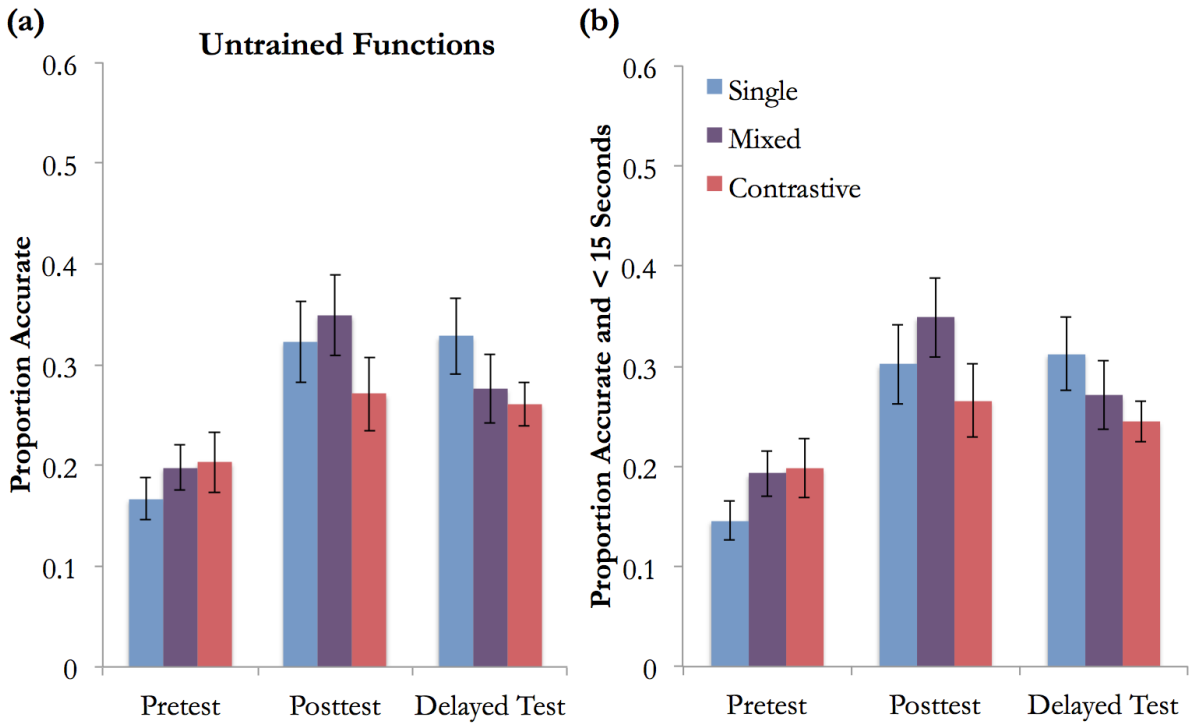


Figure 4.7. Mean (a) accuracy and (b) fluency on Untrained Functions.

Error bars are ± 1 standard error.

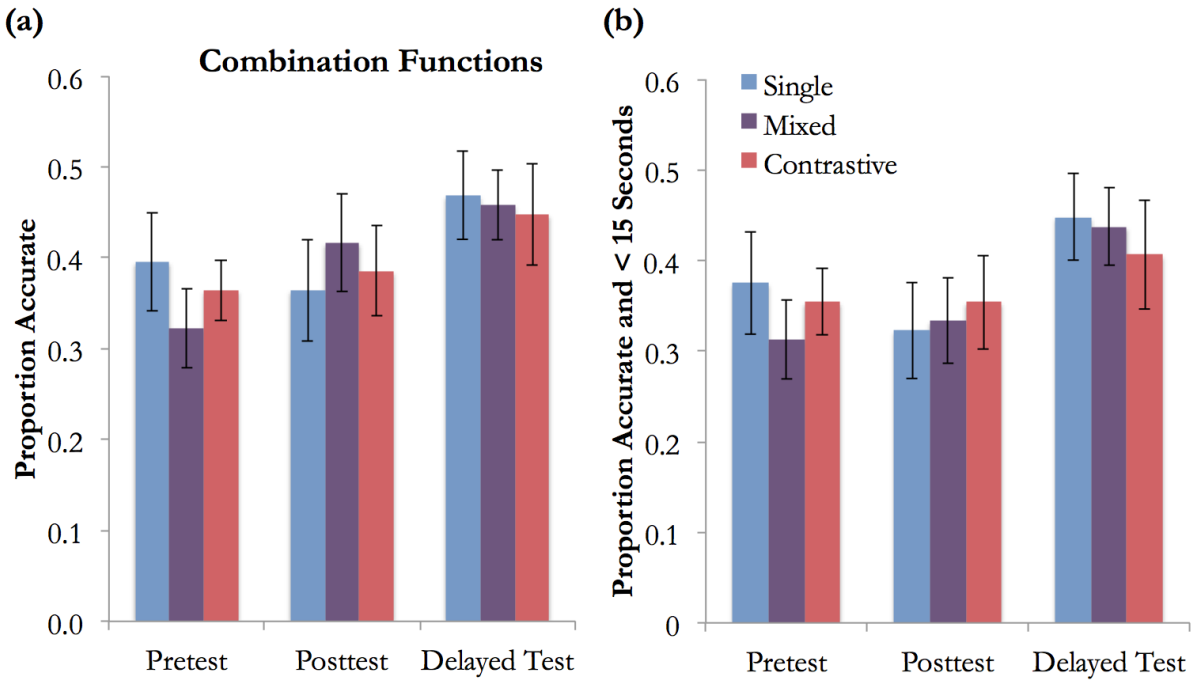


Figure 4.8. Mean (a) accuracy and (b) fluent accuracy on Combination Functions.

Error bars are ± 1 standard error.

Figure 4.8a shows the average combination functions (CF) accuracy by condition. There were no condition differences on CF, *p*'s > .10. Interestingly, the *mixed* condition was the only condition with marginal improvement from pretest to immediate posttest, $t(23) = 1.99$, $p = .06$, $d = .39$, and significant improvement from pretest to delayed test with medium effect sizes, $t(23) = 2.50$, $p = .02$, $d = .67$. *Appendix H* contains more details of these analyses.

Fluency

Figure 4.3b shows the mean fluent accuracy on all assessment items. Similar to accuracy results, the *single* condition did better than the *contrastive* condition overall with a medium effect size, $t(46) = 2.14$, $p = .04$, $d = .62$. The *mixed* condition also did marginally better than the *contrastive* condition, but the difference was of a small effect size, $t(46) = 1.66$, $p = .10$, $d = .48$.

Figures 4.4b, 4.5b, 4.7b, and 4.8b show the fluent accuracy on each assessment item type. In terms of fluent accuracy by item types, the only condition difference was on TI, when the *single* condition reliably perform better overall than the *contrastive* condition with a medium effect size, $t(46) = 2.16$, $p = .04$, $d = .62$. More details of these analyses are in *Appendix H*.

Progression of Learning

Table 4.1 shows the training performance by condition. The *mixed* group spent about 11 less minutes than the *single* group on the training, but there were no reliable condition differences on the total time spent on the training, nor on the number of learning trials, *p*'s > .10.

There was no condition differences on training accuracy, $F(2,69) = 4.14$, $p = .02$, $\eta^2_p = .11$, but there was a marginal main effect of condition on training RTc, $F(2,69) = 2.53$, $p = .09$, $\eta^2_p = .07$. *Figure 4.9* shows the average RTc by training quartiles. Both the *contrastive* and *single* groups tended to take about a second longer to get each question correctly than the *mixed* group,

$t(46) = 2.21$, $p = .03$, $d = .64$, and $t(46) = 1.85$, $p = .07$, $d = .54$, respectively. The difference between the *single* and *mixed* conditions was marginally significant, and these differences were just about 1-2 seconds long, but both had medium effect sizes. The difference between the *contrastive* group and the *mixed* group appeared in the first two quartiles into the training, $t(46) = 2.22$, $p = .03$, $d = .67$, and $t(46) = 2.34$, $p = .02$, $d = .64$, but not during the latter two quartiles. The *single* group, on the other hand, started with similar RTc as the *mixed* group but ended up taking marginally longer on the 3rd quartile and significantly longer on the 4th quartile, both with medium effect sizes, $t(46) = 1.77$, $p = .08$, $d = .51$, $t(46) = 2.10$, $p = .04$, $d = .61$, respectively. This was generally true even when we compared RTc by trial type (i.e., compare just contrastive trials of the *mixed* group with the *contrastive* group). The *contrastive* and *single* conditions did not differ on RTc on any quartiles, p's > .10.

There were no condition differences on fluent accuracies during the training, $F(2,69) = .54$, $p > .10$. Within the *mixed* condition, there were no differences in raw accuracy, RTc, nor fluent accuracy among the *contrastive* and *single* trials, p's > .10.

Condition	Total Trials	Minutes on Training	Training Accuracy	Training RTc	Training Fluency
Single	260.33 (26.13)	59.96 (6.05)	0.32 (.02)	6.39 (.49)	.30 (.02)
Contrastive	248.92 (29.56)	53.08 (10.08)	0.34 (.02)	6.66 (.52)	.32 (.02)
Mixed	211.21 (18.62)	48.83 (6.11)	0.31 (.02)	5.29 (.34)	.30 (.02)
- Single	124.96 (15.07)		0.32 (.02)	5.32 (.31)	.31 (.02)
- Contrastive	126.67 (14.88)		0.31 (.02)	5.29 (.37)	.30 (.02)

Table 4.1. Training means across conditions. Standard errors are in parentheses.

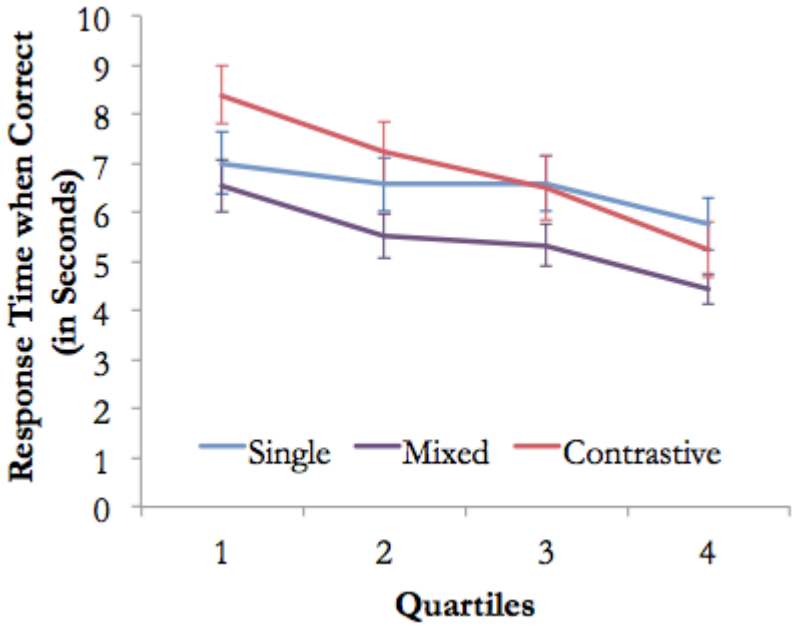


Figure 4.9. Mean RTc by training quartiles. Error bars are ± 1 standard error.

Survey Questions

There were no condition differences on any of the survey responses, *p*'s > .10. In general, participants found the training modules to be moderately enjoyable (*M* = 3.28, *SD* = .63) and helpful (*M* = 3.42, *SD* = 1.47). They reported to be quite motivated and engaged in the study (*M* = 5.17, *SD* = 1.08). On the metacognitive questions, on a scale from 1-6 (1 = "not at all", 6 = "very well"), they rated their prior knowledge of Sine and Exponential functions to be 1.93 (*SD* = .95) before the training and 3.17 (*SD* = 1.13) after the training, and 2.22 (*SD* = 1.15) at delayed test. On the same scale, we asked, "How much do you think you will remember a week from now", participants predicted that they would forget a lot of the materials learned, giving an average rating of 2.69 (*SD* = 1.13). They also judged the delayed test to be difficult (*M* = 5.22, *SD* = .91), but that they were highly motivated to do well (*M* = 5.35, *SD* = .91).

Interestingly, participants generally did not do well on description items at both immediate posttest and delayed test, in which we asked them to choose from a set of 8 possible answers, to describe the transformation shown in 4 equations. At immediate posttest, they scored an average of 35% correct (*SD* = 26%, or 1.42 questions correctly out of 4), and at delayed test 30% correct (*SD* = 23%, or 1.19 questions correctly out of 4).

DISCUSSION

Experiment 4 further confirmed the idea that a brief period of training designed to foster perceptual learning can improve learners' abilities to classify graphs of mathematical functions. Participants were better at the end of training than at the beginning, and this improvement transferred to new instances of familiar and unfamiliar function families. Transfer to new instances was, once more, evidence for the fact that learning in this task was not about memorization of exemplars during training. PALMs in general promoted better extraction of patterns and transformations that can be applied to new function families.

The three conditions did not differ on overall training efficiency, but when considered by item type, the *single* condition was more efficient than the *contrastive* condition on TI (marginal) and UF trial efficiencies. The *mixed* condition was better than the *contrastive* condition on TF/NI time efficiency. These differences had small to medium effect sizes.

In terms of accuracy, the *single* condition was reliably better than the *contrastive* condition on overall accuracy gain. This held for TI, and marginally so on Exponential TF/NI and UF accuracy gain. The *mixed* condition in some cases also did better than the *contrastive* condition, reliably so only at posttests when all items were combined and separately on TF/NI immediate posttest. There were no reliable accuracy differences between the *single* and the *mixed* conditions.

The condition differences were sparse in terms of fluency. The only reliable differences on fluent accuracy were between the *single* and the *contrastive* conditions on the overall fluent accuracy gain and on overall TI fluent accuracy gain (marginally so on average TI fluent accuracy gain). The *mixed* condition was marginally better than the *contrastive* condition on overall fluent accuracy gain.

The *single* and *mixed* conditions led to transfer to untrained functions, but the *contrastive* condition did not, and the *mixed* condition was the only one that led to transfer to combination functions. Taken together, despite the nuances in the results, we can conclude that in general, the *contrastive* condition was not as effective as the other two conditions for transfer. It was also less efficient in some cases than the *single* condition. The *mixed* condition produced similar transfer and retention as the *single* on transfer and retention, but in some cases had better time efficiency.

The disadvantage of training with only contrastive examples was apparent immediately after the training. Even though they expected to be tested on graphs without the canonical function present, the exclusive and constant use of these graphs, which contained both the canonical function and its variation, may have led participants to develop a different learning and/or performing strategy. It was possible that these participants were learning to classify graphs based almost exclusively on a comparison between the basic function *and* its variation, rather than based on aspects of the to-be-classified function by itself. That is, they may have developed a strategy of classifying $y = sin(2x)$ based on how it related to $y = sin(x)$ rather than based on other important aspects of the to-be-classified function (i.e., the point where the function intersects the axes). As a result, they became dependent on the graphical exposure of the canonical function to be able to classify its variations. This theory is consistent with participants'

self-reports[10], that they used the "reference line" during the training, but felt lost without it during the posttests. For example, one *contrastive* participant wrote "*During the training I learned how to compare the two lines and I finally noticed some sort of pattern when I compared the lines. The posttest was hard because I didn't have the dotted line to compare it to, so I felt like I was guessing a lot more."* It seems, then, that the *contrastive* graphs may be more useful when given as feedback only (*single* condition), and may make participants too dependent on them when used during the active classification task as well (*contrastive* condition).

Furthermore, the *contrastive* group was still able to improve on TI and TF/NI items, but not UF items. This finding replicated results from Silva & Kellman (1999, unpublished data), in which they provided participants with either contrastive graphs as both the to-be-classified graph and the feedback (same as our *contrastive* condition) or single graphs as the to-be-classified graph and contrastive graphs as feedback (same as our *single* condition). They found that overall, training with *contrastive* graphs fostered learning on trained functions but not to untrained functions. Even though the canonical functions were not presented in the TI and TF/NI assessment items, by the time subjects received this posttest they already had a very strongly activated representation of the studied canonical functions. Thus, they could probably rely on their memory representations of the familiar basic functions to perform in the TF/NI. This strategy could not be used in the UF because the unfamiliar canonical functions were not seen during training, and consequently subjects did not develop strong active memory representations of them.

Remarkably, the *mixed* condition was the only condition with successful transfer to all item types, including the CF items. CF items involve two function families combined in each

[10] In response to "What were your strategies during the training? Are they different from your strategies in the posttest? If so, how?"

graph, which must be compared and integrated if they are to be correctly classified. It is possible that a *mixed* practice with single and contrastive graphs allowed participants to benefit from a comparative strategy from contrastive trials and the challenge of having a mixture of trial types developed in them a more flexible representation of each transformations, in a way that the *single* condition could not.

Another advantage of the *mixed* condition was that it was equally efficient in terms of trials as the *single* condition, and was more time efficient than the *single* condition in some cases. Despite having lower accuracy during the training, the *mixed* group spent less time on each question than the other two groups. As a result, they took 11 minutes less than the *single* group, yet did equally well as the *single* group on transfer and retention. The *mixed* practice may have discouraged participants from dwelling as long on *contrastive* trials presumably because they knew they had to not be reliant on the contrastive examples. Having variations during training also may play a role in enhancing speed of processing on *single* trials. More research is needed to understand whether or not this was the case.

Interestingly, there were some intricacies in differences across conditions on Sine versus Exponential functions. The *single* condition did better than the *contrastive* condition on Sine TI (but not Exponential TI), and on Exponential TF/NI accuracy gain (but not Sine TF/NI). A closer look revealed that in general, the gains in performance on Sine were better preserved than performance on Exponential functions. Despite differences at the immediate posttest, after a week, there were no condition differences on Exponential TI and TF/NI. It is unclear why this was the case, but one possibility is that it is generally easier to recognize the transformations on Sine functions than on Exponential functions. At pretest, participants tended to show higher accuracy on Sine items (M = .21, SD = .15) than on Exponential items (M = .17, SD = .13)

though with a small effect size, $t(71) = 2.12$, $p = .04$, $d = .29$. Because the learning gains on Sine were better preserved, the advantage of *single* over *contrastive* was easily seen with Sine TI. However, perhaps because of the relative ease of Sine functions, participants generally did not find transfer to Sine TF/NI to be too difficult.

In summary, contrary to prior research showing the benefits of contrastive comparisons, we found that pure *contrastive* training was not as good as training with some or no contrastive learning trials. When given as a to-be-classified graph, *contrastive* graphs may have led participants to classify the variations mostly as compared to the canonical functions. This may have harmed their performance when having to classify graphs that did not contain the canonical functions and graphs that involved different canonical functions and more complex combination functions. On the other hand, when *contrastive* graphs were used only during feedback (*single* condition), they seemed to have fostered near transfer to both familiar and unfamiliar function families, but did not facilitate far transfer to the combination functions. *Mixed* practice provided the best of all conditions: it allowed for retention and transfer to all assessment items while doing so in an efficient manner.

GENERAL DISCUSSION

The contrastive experience was different between the two experiments (side-by-side contrast in Experiment 3, vs. overlapped in Experiment 4), yet across two very different domains, we found a consistent pattern: training with only *contrastive* comparisons can be detrimental for learning, and training with a *mixture* of contrastive and single trials was generally equivalent to training with the *single* condition.

There were however, some differences in the patterns of results between the two experiments. In Experiment 3, the *contrastive* condition did equally well as the other two conditions at immediate posttest, so that its disadvantages only occurred at delayed test. In Experiment 4, in some cases, the *contrastive* condition produced lower performance even at immediate posttest, which persisted at delayed test (except on TF/NI items). One potential explanation is based on the different nature of the *contrastive* experience across the two domains. The *contrastive* graphs may have become a crutch that when removed at the immediate posttest, participants were not able to apply what they have learned without it. In the case of ECG, participants may have simply ignored the Normal ECG altogether. If so, because participants successfully reaching learning criteria, it should be of little surprise that their performance on the immediate posttest was the same across all three training conditions, but that the exclusively contrastive experience could not produce robust learning for long-term retention.

A similar explanation may be provided for another difference between the two experiments regarding the benefit of the *mixed* condition on efficiency. In Experiment 4, the mixed condition proved to be more time efficient than the *single* condition in some cases, but it was not the case with ECG in Experiment 3. Having the *contrastive* graph overlapped with the to-be-classified graph was a helpful guide to the relevant transformation, which could potentially shorten training time. Having a Normal ECG on one side of the screen presented participants with more information to be processed, which could elongate training time without enhancing learning.

In both experiments, the *mixed* condition was not better than the *single* condition on accuracy and fluency overall. Why was this the case? One possibility is the *contrastive* environment provided was not particularly conducive to discrimination learning. The *contrastive*

display was available for participants, but it was provided passively, such that participants were never explicitly asked to compare cases in any of the experiments. Participants had an active task on all contrastive trials, but the task did not require them to look at two displays and decide, "Which is an example of category X?" Nevertheless, the results are surprising in that some published data on comparisons show learning effects in situations not dissimilar to the present procedures, without requiring a forced-choice identification (e.g., Rittle-Johnson & Star, 2009). A second possible non-optimal feature of comparisons here is that because all of the possible category labels were present, there was no guidance as to what the relevant features were that should be compared and contrasted. Whether comparison aids learning may be depend on details of how the learners are engaged with the comparison.

Some prior research has suggested that guidance toward comparisons is often needed to maximize the benefit of comparisons. For example, participants were explicitly told to think about similarities between cases (e.g., Gick & Holyoak, 1983; Gentner et al., 2003; Loewenstein, Thompson, Gentner, 1999), describe similarities and differences (Rittle-Johnson & Star, 2009), or answer a question that directly refers two cases at the same time (Rittle-Johnson & Star, 2007). There is also evidence that explicit instructions to make comparisons and contrast are useful for ECG interpretation (Ark, Brooks, & Eva, 2007). Ark et al. (2007) trained undergraduates with just 4 examples of each of 8 categories of ECG diagnoses. They provided participants with a list of key features for each diagnoses, had students practice identifying the relevant features with corrective feedback. They varied whether or not participants received explicit instructions to contrast the examples from the diagnostic category being learned with a normal ECG and with another confusable diagnostic category. Another manipulation was prior to the posttests, participants were either explicitly told to look for similarities and differences and

to use the feature list to make each diagnosis, or simply told use whatever strategy that comes naturally to them. On both the immediate and delayed tests, they found that explicit instruction to compare and contrast produced higher performance than non-contrastive training and those who were reminded to compare on the posttests did better on those who were not[11]. Similarly, in the classic analogical reasoning study using the Dunker Radiation problem (Gick & Holyoak, 1980), after studying an isomorphic problem and its solution, they were given a new, isomorphic problem to solve. Without explicit hints, participants were very unlikely to recognize that the similarities in problem structure between the two isomorphic problems. However, when they were prompted to make comparisons across the two isomorphic problems, they were able to generalize the solution to the new radiation problem (Catrambone & Holyoak, 1989; Gick & Holyoak, 1983).

The current studies cohere with these results. Comparison is not inherently good for promoting learning and transfer, rather its effects may depend on whether it supports the processing of relevant features and relations that are essential for perceptual learning. This research is consistent with other studies showing the effect of comparison being limited by the type of category to be learned (characterized by diffuse vs. localized) features; by the learners' prior knowledge; always having a contrastive experience may cause the learners to become too dependent on having the representation and their representations are fixed in a way that isn't best for transfer; by the amount of guidance provided; and the similarity of examples and how they are presented (e.g., Lee, Betts, Anderson, 2015). These findings are important, suggesting that it

[11] The feature list containing the key features for each diagnosis was also available for participants at testing. These findings would have provided support for the importance of comparisons in ECG interpretation learning, but the contrastive training condition had more time in the training than those in the non-contrastive condition so time differences in the training confounded their findings. Furthermore, it is unclear how much of what was learned was simply due to memorization versus perceptual learning because there were few instances during the training, and participants performed much better on previously seen ECG than new ECG at delayed test.

is not the case that *any* contrastive presentation structure encourages facilitates discovery of relevant features or patterns. Rather, having either too many opportunities for comparison or too much information available for comparison may actually be detrimental. The results also support the interpretation that comparison is not a learning mechanism, but rather a procedure (or, more correctly, a variety of procedures), which can sometimes facilitate perceptual learning (Kellman, Massey & Son, 2010).

Further Questions

We can relate these findings to those of the Experiments 1 and 2, which showed that studying passive presentations was not enough for learning, but that a combination of early passive study and active classification practice was more beneficial. This suggests that perhaps a more gradual fading method from *contrastive* to *single* would have worked better for the *mixed* condition, where the *contrastive* experience occurs first, followed by the *single* condition. This may be particularly effective in the case of graphical transformation. The initial *contrastive* condition may guide participants to the differences between the two functions for identifying the transformation.

It is possible that our results would have been improved by additional procedures that directed attention toward the comparison making process. Prior research has often involved additional efforts to optimize the benefit of comparisons. We may be able to maximize the benefit of comparisons when comparison opportunities are provided more selectively, and in a way that we can more directly (even implicitly) guide participants' attention to the relevant features. One possibility is to present the comparisons adaptively, only when they are needed, and to modify the question-answer format so that participants must directly engage in the

comparison process in order to answer the question. We explored this possibility in the next two experiments.

CONCLUSION

These experiments showed that training with only contrastive opportunities can be detrimental for learning and retention, but left open the possibility that some contrastive opportunities can be helpful for transfer and efficiency.

CHAPTER 6

Adaptive Comparison

INTRODUCTION

Comparisons can be useful, but what *kind* of comparison is useful? *When*? And for *whom*? In much of learning literature, the benefits of comparisons are found when the same comparisons apply to all learners.

Not all comparisons share the same benefits. Comparing multiple members of the same category (i.e., within-category comparison) provides information about category characteristics, while contrasting a category member against a non-category member (i.e., between-category comparison, or "contrast") provides information about "category boundaries", or the information about what distinguishes category members from non-members. Indeed, Hammer, Hertz, Hochstein, and Weinshall (2009) found different benefits when comparisons are made across two exemplars from the same category versus across two exemplars from two different categories. Between-category comparisons are often very helpful in that they may decisively indicate the relevant dimensions. This is especially true for minimal contrasts. When two different-category exemplars are similar in most of their properties, the only differentiating dimension must be relevant for the classification task. On the other hand, comparisons of same-category exemplars can reveal the irrelevant dimensions. When two same-category exemplars are similar in some of their properties, the characteristic differing between the two exemplars are necessarily irrelevant (Hammer et al., 2009; Gentner & Markman, 1994). We cannot make judgments on the remaining shared features, however, because they may or may not be relevant.

Many studies have explored the benefit of within-category and between-category

comparisons separately (e.g., Kotovsky & Gentner, 1996; Namy & Gentner, 2002; Andrews et al., 2005; Hampton, Estes, & Simmons, 2005; Kalish & Lawson, 2007; Kurtz & Boukrina, 2004), but only a few have shown success with combining between-category and within-category comparison opportunities in the training (e.g., Ankowski, Vlach, & Sandhofer, 2012; Weitnauer, Carvalho, Goldstone, & Ritter, 2014; Hammer et al., 2009). Both types are comparisons are important; when both within and between-category comparisons are presented appropriately, they have the potential to optimize the extraction of features and relations that define a category and distinguish it from others. In Experiments 5 and 6, we examined one way to harness adaptive methods to show between- and within-category comparisons when learners need them most.

How do we know when learners need which type of comparison? One of the few examples we could find that adapted comparison training to the individual was from Siegel & Misselt (1984). They gave participants adaptive feedback and discrimination training based on the type of mistakes each learner made while identifying English-Japanese word pairs. On each trial, learners were asked to type the Japanese word that corresponded to a given English word. When they typed a response that *was not* an answer to another item in the drill list, the adaptive feedback showed a message with the correct answer. When the learner typed a response that *was* an answer to another item in the drill list, they called this "discrimination error", and the adaptive feedback consisted of both the answer to the word in question and the English correspondence for the word that they typed.

This "discrimination error" could also trigger follow-up discrimination training. In the discrimination training, the missed item and item with which it was confused were presented simultaneously, to allow for comparison of their similarities and differences. Siegel & Misselt

found that this adaptive feedback with discrimination training was significantly more effective at helping students memorize the word pairs than other forms of feedback studied (not adaptive, or without discrimination training).

Though this study addressed the memorization of word pairs, the potential benefits of adapting comparisons and discrimination training are present in the context of perceptual category learning. In this study, rather than presenting a paired-comparison after *each* error, we distinguished between two forms of classification errors by looking at participants' *pattern* of error, and triggered either between-category comparison or within-category comparison based on those error patterns.

One of those patterns was similar to what Siegel & Misselt called "discrimination error"; when learners chose category B twice in a row when the answer was category A, we inferred that participants may have had trouble detecting the relevant characteristics that distinguish between those two categories. They were then presented with between-category discrimination practice between instances of A and B to highlight those relevant dimensions. The other pattern of error was when participants chose an instance of category B and another of C on two separate trials when the answer was A. From this, we inferred that they have had trouble seeing what constituted category A, so we presented within-category comparisons of two A instances to highlight the relevant dimensions of A and its irrelevant variations.

In Experiments 5 and 6, we asked, how can we tailor paired-comparisons to learners' need to maximize training efficiency? We used each learner's pattern of error to trigger either between-category comparisons or within-category comparisons. The goal was to test whether adaptive comparison can improve training efficiency without sacrificing transfer and retention. In both experiments, we distinguished between two types of trials within each PALM: the active

classification trials (*adaptive learning, AL*) seen in previous experiments, and *comparison trials,* referring to the between- and within-category comparisons triggered by participants' responses. As a control for the mere presence of comparison trials, we also tested a group who received comparisons about equally often as those in the adaptive comparison condition, but where the comparisons were not triggered by or related to error patterns.

Experiment 5

METHOD

Participants

108 undergraduates (mean age = 22.33; 70 female) without prior knowledge of ECG interpretation from University of California, Los Angeles participated for course credit.

Design

Participants were randomly assigned into one of three training conditions: (1) baseline adaptive learning with classification trials only (*AL*), (2) adaptive learning with classification trials and adaptive comparisons (*AL/AC*), (3) adaptive learning with classification trials and non-adaptive comparisons (*AL/NC*).

Materials

The materials were the same as those used in Experiment 3.

We added two types of *comparison trials*: between-category comparison (AB trials) and within-category comparison (AA trials). The AB displays consisted of a side-by-side representation of two ECG halves from (the same side of) two different categories of heart patterns. The within-category (AA) displays consisted of a side-by-side representation of two ECG halves from (the same side of) the same category of heart patterns. Thus, one of those two

halves may have contained no relevant information for a particular heart pattern. For example, in a AB trial comparing between LAD and Anterior STEMI (LAD being the target category), participants could have seen the left-hand tracing, which were relevant to LAD on one side, and the left-hand tracings of Anterior STEMI on the other (relevant tracings for Anterior STEMI are the ones on the right-hand side).

Procedure

The procedure was identical to that used in Experiment 3. The *AL* PALM was identical to that of the *single* PALM from Experiment 3.

In the *AL/AC* PALM, participants received the same adaptive learning paradigm with classification trials as those in the *AL* PALM, with the addition of comparison trials triggered by their error patterns. When participants made a consistent miss, by choosing 2 out of 3 instances one wrong category label (e.g., both times chose Inferior STEMI when the answer was Anterior STEMI), they were given an AB trial, on which Inferior STEMI and Anterior STEMI were presented side-by-side. When instead, participants chose two different categories when the answer was another (e.g., chose Inferior STEMI and RBBB on two different instances when the answer for both was Anterior STEMI), they were presented with an AA trial in which two Anterior STEMI halves were presented on the screen. The PALM kept track of the past 3 instances of each category, so patterns with no or one intervening correct response could have triggered a comparison trial. When presented with an AB or AA trial, participants were asked to pick all of the applicable ECG halves that show a given heart pattern (e.g., "Anterior STEMI is the main diagnoses of which of these ECG traces") and were provided with three answer choices: Left, Right and Both. Thus, in AB trials, "Left" or "Right" was the correct answer. In AA trials, "Both" was correct. *Figure 5.1* shows a sample AA trial and *Figure 5.2* shows the

feedback screen of a sample AB trial. Participants returned to an AL trial immediately after each comparison trial.

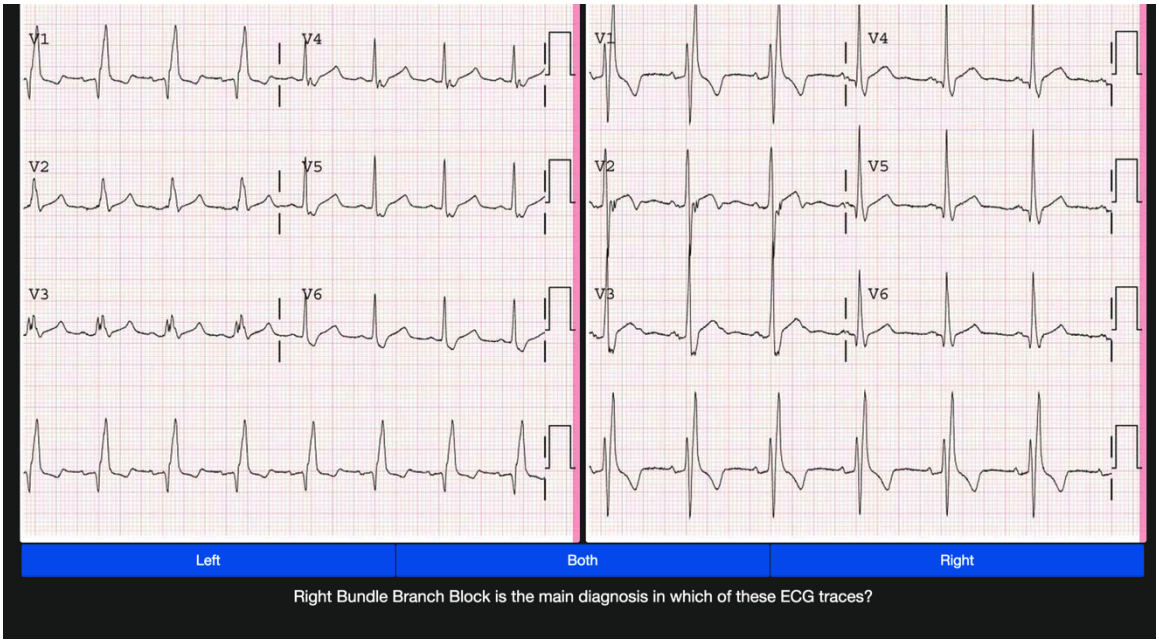


Figure 5.1. A sample AA trial.

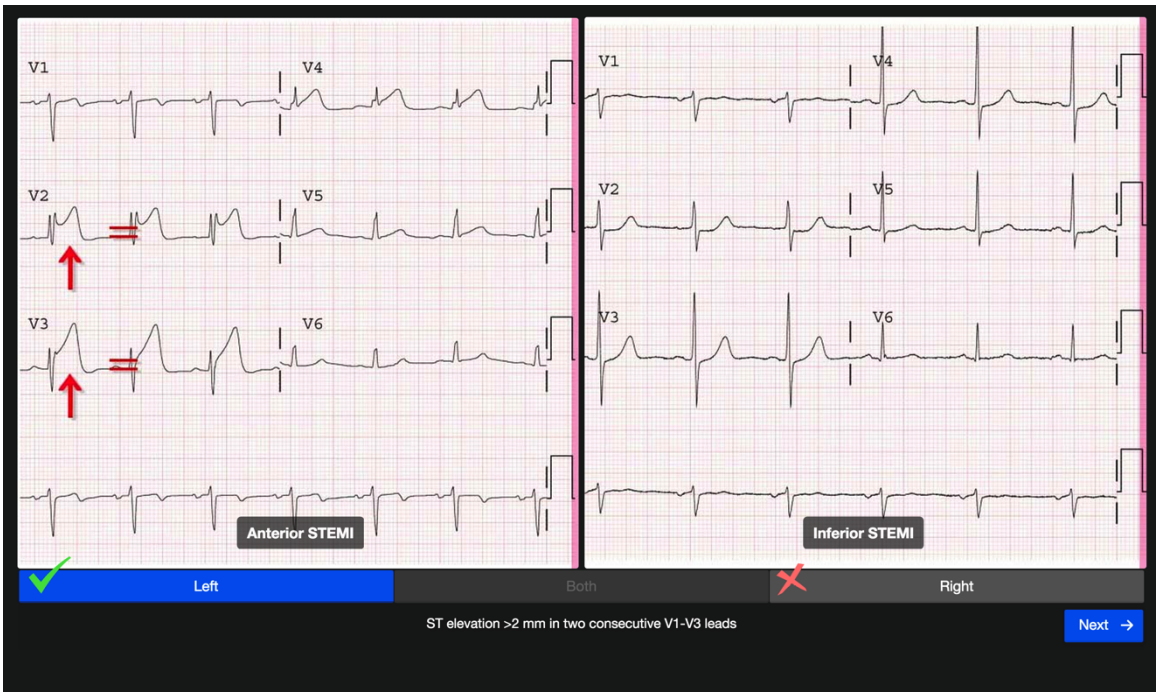


Figure 5.2. Sample feedback screen of an AB trial.

The *AL/NC* PALM group received the same adaptive learning paradigm with classification trials as well as comparison trials that were triggered by a pattern of error. Crucially, however, the type of comparison trials and the categories presented for comparison were selected at random. Thus, it was possible that after a classification trial on RBBB, participants could have received an AB trial showing Anterior STEMI and Left Axis Deviation. The total number of comparison trials and the ratio of AB and AA trials (30:70) were matched between the *AL/AC* and *AL/NC* conditions. The ratio of AB and AA trials was determined with a pilot group of *AL/AC* participants.

All three conditions used the same learning criteria as used in the previous experiments. Performance on comparison trials did not count toward retirement.

Overview of Analyses and Expected Results

The three groups did not differ on quiz performance or any survey measures. Different training conditions resulted in different rates of completion. The comparison conditions, particularly the *AL/NC*, were much more difficult for participants to complete within the time allotted than the *AL* condition. Only 55% of those in the *AL/NC* condition compared to 83% in the *AL* condition and 69% in the *AL/AC* condition completed their assigned PALMs. Similar to prior experiments, we stopped data collection when we had equal number of participants who reached learning criteria in each condition. In the following section, we report analyses for participants who completed the PALMs, and report the results from all participants in *Appendix I.1*. The same general patterns of results were found with both samples. All statistical assumptions were met.

Based on prior studies, we expected all PALMs to produce robust classification learning and transfer to new ECGs. We were also interested in whether comparison trials can improve

training efficiency without sacrificing learning and retention. To compare trial efficiency, we took into account the total number of training trials completed, including both AL and comparison trials (efficiency = posttest gains in accuracy divided by the total training trials invested). An important comparison of efficiency measures was between the *AL/AC* condition and the baseline *AL* condition. Because each comparison trial was considered a learning trial, it was possible that the *AL/AC* condition would require more trials (and time) to complete the module than those in the *AL* condition. However, it may also be possible that because *adaptive* comparisons directly targeted what the learners need, learners may acquire mastery levels faster than those in the other conditions and produce better efficiency than the *AL* condition. Because the *AL/NC* condition did not tailor comparisons to each learner's error pattern, we expected this group to require more trials (or time) than in the *AL/AC* condition to learn the discriminations needed for mastery because they could waste trials (or time) reviewing what they have recently learned. Thus, we expected the *AL/AC* condition to be more efficient than the *AL/NC* condition.

In terms of accuracy and fluent accuracy, if the *AL/AC* condition could indeed target the discriminations that learners have difficulty with, participants would learn the transformations sooner than the *AL/NC* (and perhaps also the *AL*) condition; they would develop greater fluency in processing those relations and perhaps pick up on even higher-order relations, allowing them to perform better on transfer measures. Furthermore, if any AB and AA comparison is good for learning relevant relations, we would find the *AL/NC* condition to outperform the *AL* condition on transfer.

RESULTS

Efficiency

Efficiency by Trial

Figure 5.3a shows the efficiency by *total* trials by condition. A 2 phase (pre-post, pre-delayed) x condition ANCOVA on trial efficiency with pretest accuracy as the covariate confirmed a main effect of condition, $F(2, 68) = 2.84$, $p = .07$, $\eta^2_p = .08$. After controlling for the effect of the pretest, the *AL/AC* condition ($M = .003$, $SD = .001$) produced overall higher trial efficiency than the *AL/NC* condition with a medium effect size ($M = .002$, $SD = .001$), $t(46) = 2.47$, $p = .02$, $d = .70$, and marginally higher than the *AL* condition with a small effect size ($M = .002$, $SD = .001$), $t(46) = 1.74$, $p = .09$, $d = .48$. The *AL/NC* and *AL* conditions did not differ on the overall trial efficiency, $t(46) < 1$, $p > .20$.

The benefit of the *AL/AC* was most apparent at delayed test, when *AL/AC* had higher pre-delayed trial efficiency than *AL* with a medium effect size (.0023 vs. .0013, $t(46) = 2.28$, $p = .03$, $d = .67$). There was no difference between these two conditions on pre-post trial efficiency, $t(46) < 1$, $p > .20$. *AL/AC* had higher trial efficiencies than *AL/NC* on both pre-post (.0030 vs. .0021, $t(46) = 2.17$, $p = .04$, $d = .63$) and pre-delayed efficiencies (.0023 vs. .0014, $t(46) = 2.30$, $p = .03$, $d = .67$[12]) with medium effect sizes. *AL/NC* did not differ from *AL* at either test phases, $t(46) < 1$, p's $> .20$.

There was a significant main effect of phase, $F(1, 68) = 13.44$, $p < .001$, $\eta^2_p = .17$. The immediate posttest trial efficiency ($M = .003$, $SD = .002$) was significantly higher than the delayed test trial efficiency ($M = .002$, $SD = .001$). There was also a main effect of pretest, $F(1, 68) = 15.99$, $p < .001$, $\eta^2_p = .19$. The pretest accuracy negative correlated with the immediate

[12] These differences between *AL/AC* and *AL/NC* were not reliable when we ignore pretest variations, assuming that they were due to chance. However, pretest may not have been entirely random, considering that the *AL/NC* condition was considerably harder for participants to finish in time, and *AL/NC* had slightly higher pretest accuracy, suggesting that those who did may have been higher performing.

posttest trial efficiency, $r(72) = -.47$, $p < .001$, and the delayed test trial efficiency, $r(72) = -.35$, $p = .003$. There were no significant phase x pretest and phase x condition interactions, p's > .10.

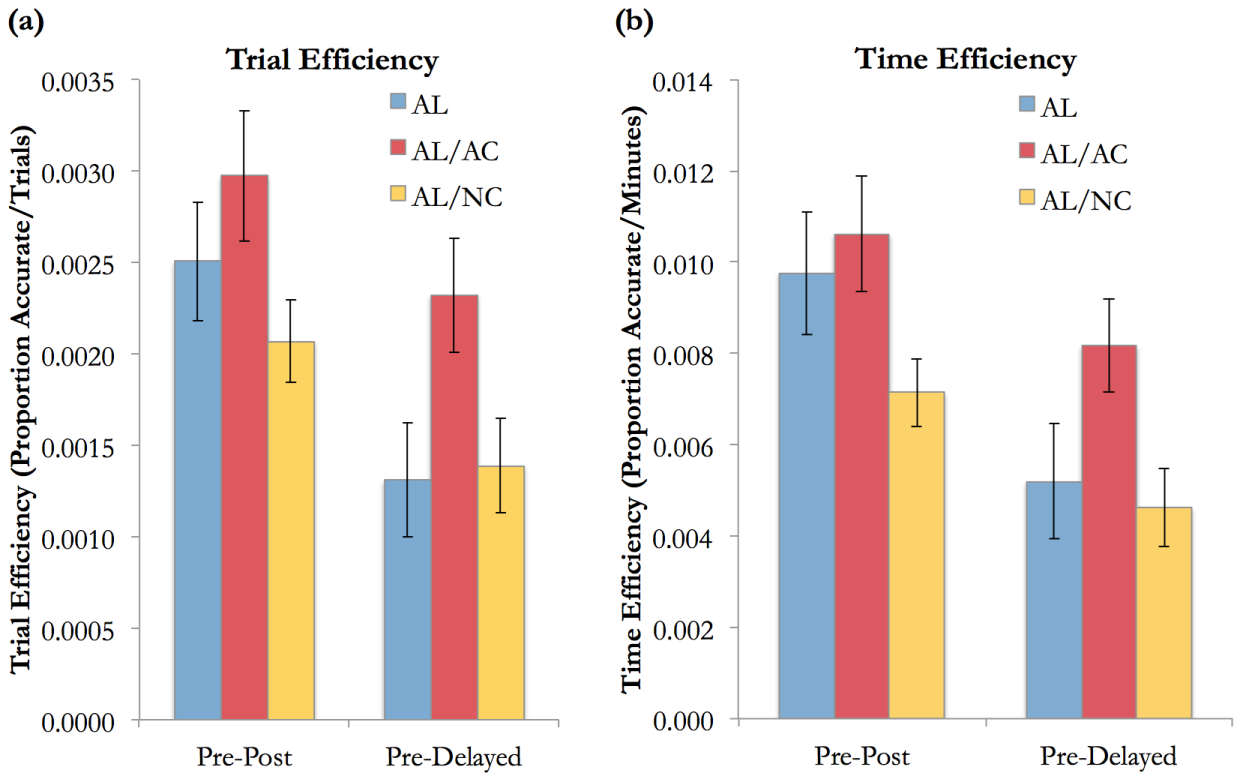


Figure 5.3. Mean efficiency (a) by trial and (b) by time.

Efficiency by Time

Figure 5.3b shows the time efficiency by condition. Time efficiency demonstrated a similar pattern as trial efficiency. There was higher overall time efficiency with *AL/AC* than *AL/NC* with large effect size (.009 vs. .006), $t(46) = 2.84$, $p = .007$, $d = .82$. *AL/AC* did better than *AL/NC* on both pre-post, $t(46) = 2.36$, $p = .02$, $d = .69$ and pre-delayed time efficiency, $t(46) = 2.65$, $p = .01$, $d = .76$. *AL/AC* and *AL* did not differ on overall time efficiency, $t(46) = 1.21$, $p > .20$, but *AL/AC* was marginally superior than *AL* on pre-delayed time efficiency with a medium

effect size (.0082 vs. 0052, $t(46) = 1.84$, $p = .07$, $d = .54$[13], but not on the pre-post trial efficiency, $t(46) < 1$, $p > .20$). *AL* did not differ from *AL/NC* on overall time efficiency, $t(46) = 1.12$, $p > .20$, but did marginally better than *AL/NC* on pre-post time efficiency, $t(46) = 1.71$, $p = .10$, $d = .49$, and not on pre-delayed time efficiency, $t(46) < 1$, $p > .20$.

Accuracy

Figure 5.4a shows the mean accuracy on the assessments by condition. Accuracy at pretest was highest for the *AL/NC* group (34%, compared to 28% from the other two groups). This suggested that those who reached mastery criteria with this module were generally higher performing. This confirmed the difficulty of mastering the training module with non-adaptive comparisons. These condition differences at pretest, however, were not statistically reliable (*p*'s > .10).

Participants from all three conditions generally showed strong learning gains and retention, $F(2, 138) = 145.85$, $p < .001$, $\eta^2_p = .68$, from pretest ($M = .30$, $SD = .15$) to immediate posttest ($M = .67$, $SD = .13$), $t(71) = 15.86$, $p < .001$, $d = 2.68$, and from pretest to delayed test ($M = .54$, $SD = .17$), $t(71) = 10.25$, $p < .001$, $d = 1.52$, with very large effect sizes. Accuracy also dropped significantly from immediate posttest to delayed test, $t(71) = 6.61$, $p < .001$, $d = .85$.

There was a statistically significant main effect of condition, $F(2, 69) = 3.67$, $p = .03$, $\eta^2_p = .10$. Both the *AL/AC* condition ($M = .52$, $SD = .10$) and the *AL/NC* condition ($M = .53$, $SD = .09$) produced significantly higher overall accuracy than the *AL* condition ($M = .46$, $SD = .11$), $t(46) = 2.15$, $p = .04$, $d = .62$, and $t(46) = 2.44$, $p = .02$, $d = .70$, respectively, with medium effect sizes. The *AL/AC* and *AL/NC* conditions did not reliably differ in overall accuracy, $t(46) < 1$, $p > .20$. There was no significant phase x condition interaction, *p*'s > .10.

[13] This difference between the *AL/AC* and *AL* conditions was not reliable when we ignored pretest variations.

Planned comparisons showed that the condition differences were only robust at delayed test, but not at immediate posttest ($t(46) < 1.4$, p's $> .16$. At delayed test, *AL/AC* and *AL/NC* did significantly better than *AL* (.59 and .57 vs. .46, $t(46) = 2.66$, $p = .01$, $d = .77$, and $t(46) = 2.34$, $p = .02$, $d = .68$, respectively).

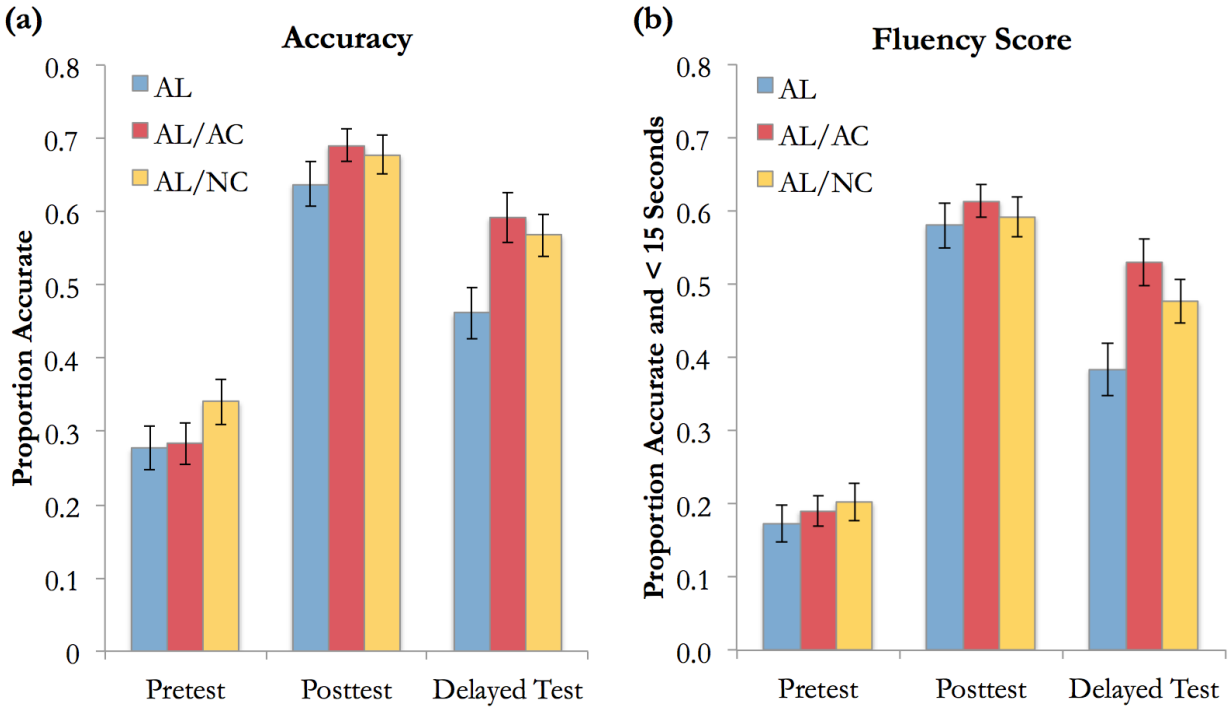


Figure 5.4. Mean (a) accuracy and (b) fluent accuracy.

Accuracy Gain

Analyses with accuracy gain confirmed that the *AL/AC* condition produced marginally higher overall accuracy gains than the *AL* condition (.36 vs. .27, $t(46) = 1.72$, $p = .09$, $d = .50$). This difference was robust only on pre-delayed accuracy gain (.31 vs. .18, respectively), $t(46) = 2.34$, $p = .02$, $d = .68$, and not on pre-post accuracy gain, $t(46) < 1$, $p > .20$. *AL/AC* had numerically higher accuracy gain than the *AL/NC* condition overall and particularly on pre-

delayed gain, but this was not a reliable effect on overall gain, $t(46) = 1.52$, $p = .14$, on pre-post gain, $t(46) < 1$, $p > .20$, nor on pre-delayed gain, $t(46) = 1.45$, $p = .15$.

Fluency

Figure 5.4b shows the mean fluent accuracy by condition. The patterns of fluent accuracy were similar to that of raw accuracy. The 3 phase x 3 condition ANOVA confirmed a marginal main effect of condition on overall fluency score, $F(2, 69) = 3.04$, $p = .06$, $\eta^2_p = .08$. The *AL/AC* condition produced higher overall fluency score ($M = .44$, $SD = .08$) than the *AL* condition ($M = .38$, $SD = .11$), $t(46) = 2.32$, $p = .03$, $d = .67$. The *AL/NC* condition ($M = .42$, $SD = .09$) also scored higher than the *AL*, but the difference was not statistically reliable, $t(46) = 1.55$, $p = .13$. There was no difference between the two comparison conditions, $t(46) < 1$, $p > .20$.

There was also a marginal interaction of phase x condition on fluent accuracy, $F(4, 138) = 1.98$, $p = .10$, $\eta^2_p = .05$. At pretest and immediate posttest, there were no differences across conditions ($t(46) < 1$, p's $> .20$), but notably, at delayed test, the *AL/AC* condition had significantly higher and the *AL/NC* condition had marginally higher fluent accuracy than the *AL* condition with large and medium effect sizes (.53 and .48 vs. .38, $t(46) = 2.99$, $p = .004$, $d = .86$, and $t(46) = 1.97$, $p = .06$, $d = .57$, respectively). *AL*/*AC* and *AL*/*NC* did not differ on their delayed test fluent accuracy, $t(46) = 1.21$, $p > .20$.

In terms of fluent accuracy gain, the only notable condition difference was that *AL/AC* did marginally better than the *AL* condition on overall fluent accuracy gain with a medium effect size (.38 vs. .31, $t(46) = 1.75$, $p = .08$, $d = .51$) and also on pre-delayed fluency gain (.44 vs. .38, $t(46) = 2.32$, $p = .03$, $d = .67$). They did not differ on pre-post fluency gain, $t(46) < 1$, $p > .20$, and there were no other reliable condition differences, $t(46) < 1$, $p > .20$.

Progression of Learning

Table 5.1 contains the descriptive statistics from the training for each condition. There were no condition differences in the total number of trials spent, the total amount of time spent, nor on RTc during the training, *p*'s > .10. However, there were significant differences among conditions on the total number of AL trials needed to reach learning criteria, $F(2, 69) = 5.42$, $p = .007$. Both the *AL/AC* and *AL/NC* conditions required fewer AL classification trials to reach learning criteria than the *AL* condition (119.54 and 131.42 vs. 154.46 trials, respectively) with large and medium effect sizes, $t(46) = 3.15$, $p = .003$, $d = .91$ and $t(46) = 2.04$, $p = .047$, $d = .59$, respectively. There were no differences between the *AL/AC* and *AL/NC* conditions on the number of comparison trials completed, $p > .10$.

Figure 5.5a shows the mean accuracy on AL trials by quartiles in the training. In terms of accuracy, the *AL/AC* group had marginally higher AL accuracy during the training than the *AL* condition with a medium effect size, $t(46) = 1.87$, $p = .07$, $d = .54$. This held only during the first training quartiles of AL trials, $t(46) = 1.93$, $p = .06$, $d = .56$, which was also a marginal effect. *AL/NC* also had numerically higher overall AL accuracy than *AL*, but the difference was not statistically reliable (.57 vs. .53, $t(46) = 1.53$, $p = .13$). There was no difference between the *AL/AC* and *AL/NC* conditions, $t(46) < 1$, $p > .20$.

Figure 5.5b shows the mean fluent accuracy on AL trials by quartiles in the training. In terms of fluent accuracy on AL trials, the *AL/AC* did significantly better and *AL/NC* did marginally better than *AL* with medium effect sizes, $t(46) = 2.46$, $p = .02$, $d = .71$, and $t(46) = 1.87$, $p = .07$, $d = .54$, respectively. The difference between *AL/AC* and *AL* was marginally reliable only on the 1st training quartile, $t(46) = 1.93$, $p = .06$, $d = .56$, and not on later quartiles, *p*'s > .10. *AL/NC* had higher fluent accuracy than *AL* on the 2nd and 3rd quartiles, $t(46) > 1.71$, $p =$

.06 to .09, d = .50 to .56, but not on the 1st or 4th quartiles, p's > .10. There were no differences between the *AL/AC* and *AL/NC* conditions on AA and AB comparison accuracies, $t(46) < 1$, $p > .20$ and $t(46) = 1.37$, $p = .18$. *Appendix I.2* contains more details of these analyses.

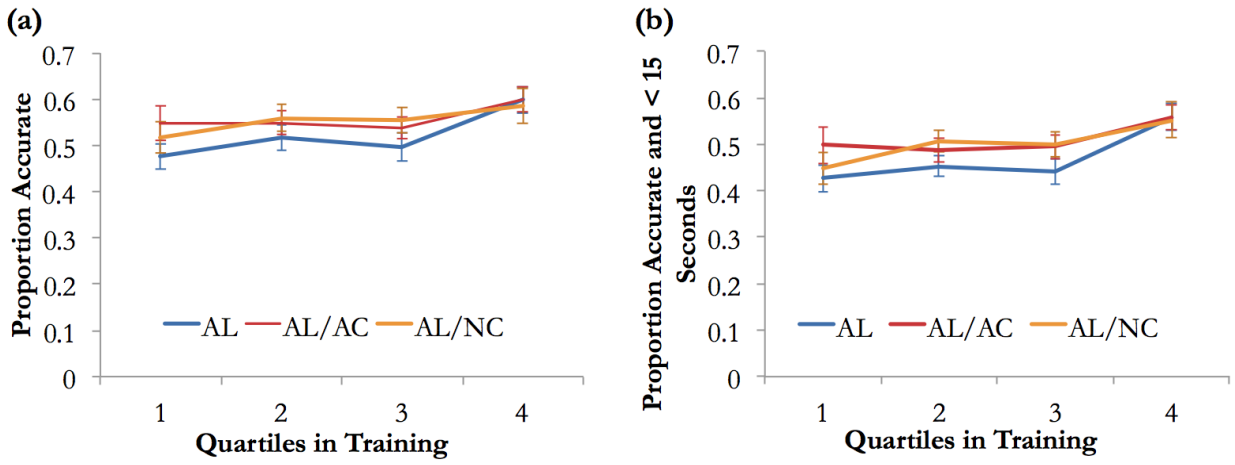


Figure 5.5. Mean (a) accuracy and (b) fluent accuracy on AL trials by quartiles in the training.

Condition	Minutes on Training	AL trials	Comparison trials	AA trials	AB trials	AL accuracy	AA accuracy	AB accuracy
AL	40.67 (2.72)	154.46 (8.73)	--	--	--	.53 (.02)		
AL/AC	42.50 (2.80)	119.54 (6.86)	36.79 (4.31)	25.71 (3.24)	11.08 (1.31)	.58 (.02)	.71 (.03)	.52 (.04)
AL/NC	47.33 (2.91)	131.42 (7.16)	41.38 (4.94)	27.79 (3.22)	13.58 (1.96)	.57 (.02)	.72 (.03)	.62 (.05)

Table 5.1. Training means in Experiment 5. Standard errors are in parentheses.

Survey Responses

There were no group differences on any survey questions, *p*'s > .10. On a scale from 1-6 (1 = not at all, 6 = very much), participants self-reported to be engaged ($M = 4.79$, $SD = .98$), found the modules to be fairly enjoyable ($M = 3.94$, $SD = 1.21$) and quite helpful ($M = 5.29$, $SD = .80$). They also thought they learned a lot from the training ($M = 5.22$, $SD = .84$).

DISCUSSION

As expected, all three training conditions produced strong learning gains, transfer and retention. Training with adaptively triggered paired-comparisons boosted trial and time efficiency. The *AL/AC* condition also performed better than *AL/NC* on overall accuracy, and better than *AL* at delayed test, suggesting that training with adaptively triggered paired-comparisons in general enhanced learning and retention. There was no difference between *AL/NC* and *AL* in terms of accuracy, but *AL/NC* had better fluent accuracy at the delayed test than *AL*. This was true for fluency on AL trials during the training as well, suggesting that training with comparisons may in general be beneficial. Taken together, training with adaptive comparisons proved to be the most effective for producing overall learning gains while being efficient.

The effect of comparison in this experiments differed slightly from that in Experiment 3, which showed the downfall of training with only *contrastive* instances. In Experiment 3, we showed participants 2 ECG halves and asked them to diagnose one of them with 7 answer choices. The *AL/AC* and *AL/NC* comparisons differed from this *contrastive* experience in two important ways. First, there were fewer opportunities for paired-comparisons; second, we provided participants with a diagnostic category label, and asked participants to choose the ECG halves that belonged to that diagnostic category. In doing so, rather than leaving it up to the

participants to figure out what features to compare, we implicitly guided participants to look for the relevant features that were specific to the target diagnostic category. This seemed to be a successful approach. The comparison trials in this experiments (even when they were not adaptive to learner error) elevated fluent accuracy during the training and transfer at the delayed test, above that of the baseline adaptive classification learning condition. This suggested that the combination of between- and within-category comparisons proved to be a potent ingredient for an effective training.

Another important finding of this experiment was that the benefits of comparison do not rest merely on some set combination of between- and within-category comparisons. Even with a similar total number of comparison trials with the same ratio of between- and within-category comparisons, the *AL/NC* condition was less efficient than the *AL/AC* condition. Furthermore, it was particularly difficult for participants to reach learning criteria in this condition, confirming the futility of the non-adaptive comparison trials. By tailoring the comparison experiences to participants' patterns of errors, we were able to better target their misunderstandings to result in enhanced learning and retention. The *AL/AC* group achieved the highest learning gains and retention, while doing so with the good efficiency, even when measured with the total number of trials completed in the training. It was possible that when customized to learners' needs, the comparison experiences guided their attention to the relevant features important for classification, and with practice with within-category comparisons, they learned to extract the relevant features for each category amidst the feature variations to develop a more flexible structural representation for each category.

In the *AL/AC* condition, many more AA trials were triggered than AB trials, suggesting that most of the time, participants were not confused between two categories, rather, they had

trouble knowing what determined one. As the training progressed, they experienced more consistent errors, which triggered more AB trials. Performance on adaptive comparison trials did not count toward module completion, but their presence enhanced performance on *AL* trials.

To the best of our knowledge, this was the first study that has attempted to adaptively trigger within- and between-category comparisons based on learner error. There are potentially better ways to adapt the training to learner error, but this experiment opened an exciting avenue that adaptive comparison together with adaptive active classification practice can enhance the effectiveness and efficiency of learning in general and PALMs in particular. To examine the generality of this finding, we looked to replicate these benefits of adaptive comparison training with the mathematics domains.

Experiment 6

In this experiment, we asked, do comparisons enhance perceptual learning of mathematical transformations? Do adaptive comparison support better training efficiency?

METHOD

Participants

72 participants (40 female, mean age = 32.25) who have passed Algebra 2 or an equivalent course from Amazon Mechanical Turk completed the study. *Appendix J* contains more demographic data.

Design

Participants were randomly assigned into one of three training conditions: (1) baseline adaptive learning with classification trials only (*AL*), (2) adaptive learning with classification

trials and adaptive comparisons (*AL/AC*), (3) adaptive learning with classification trials and non-adaptive comparisons (*AL/NC*).

Materials

All materials were identical to Experiment 4. The *AL* condition was the *single* condition from Experiment 3. We added to it between-category comparison (AB) and within-category comparison (AA) trials.

AB displays presented two graphs shown side-by-side, showing two *single* graphs from two different transformation subcategories. Recall that there were 4 categories of transformations, each with two subcategories indicating the direction of the transformation. *Figure 6.1* shows a sample AB trial in which the target function is $y = exp(x) - 2$ from the *y*-shifting downward, and the other function is $y = exp(x) + 2$ from the *y*-shifting upward subcategory. AA displays also contained two graphs presented side-by-side, except that the two plotted functions belong to the same transformation subcategory. *Figure 6.2* shows a sample AA trial in which the two main functions are $y = sin(x + 10)$ and $y = sin(x + 6)$, both from the *x*-shifting leftward subcategory. Each comparison trial provided an equation, "Which graph best matches this equation", and two answer choices "Left" and "Right" underneath each graph. After each response, the graphs were replaced with their contrastive versions, on which the gray, dotted canonical function was added.

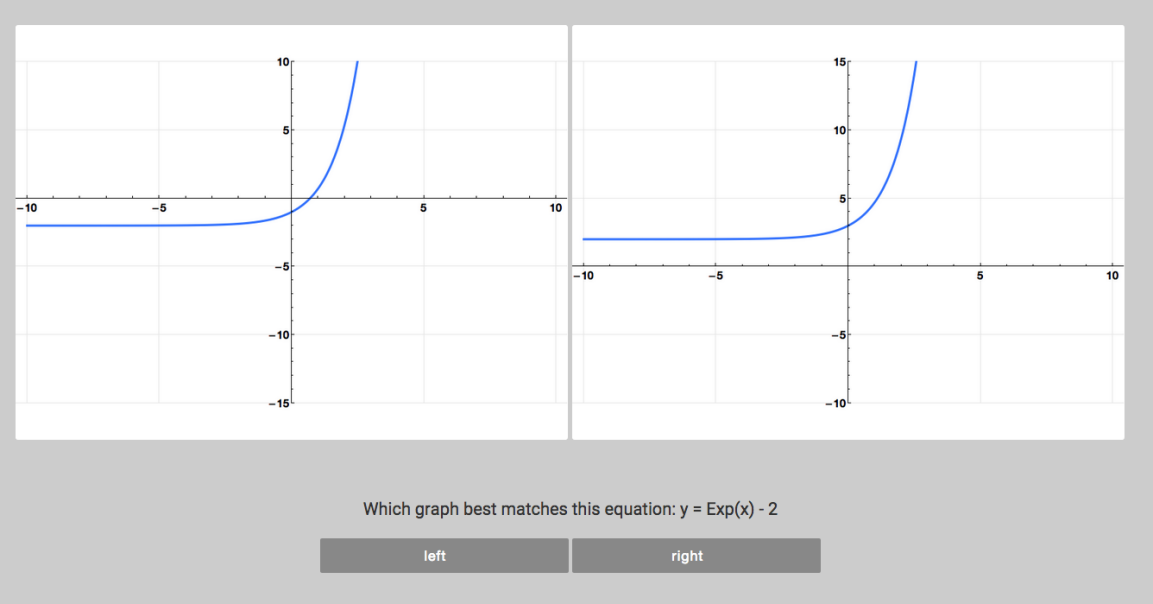


Figure 6.1. A sample AB trial

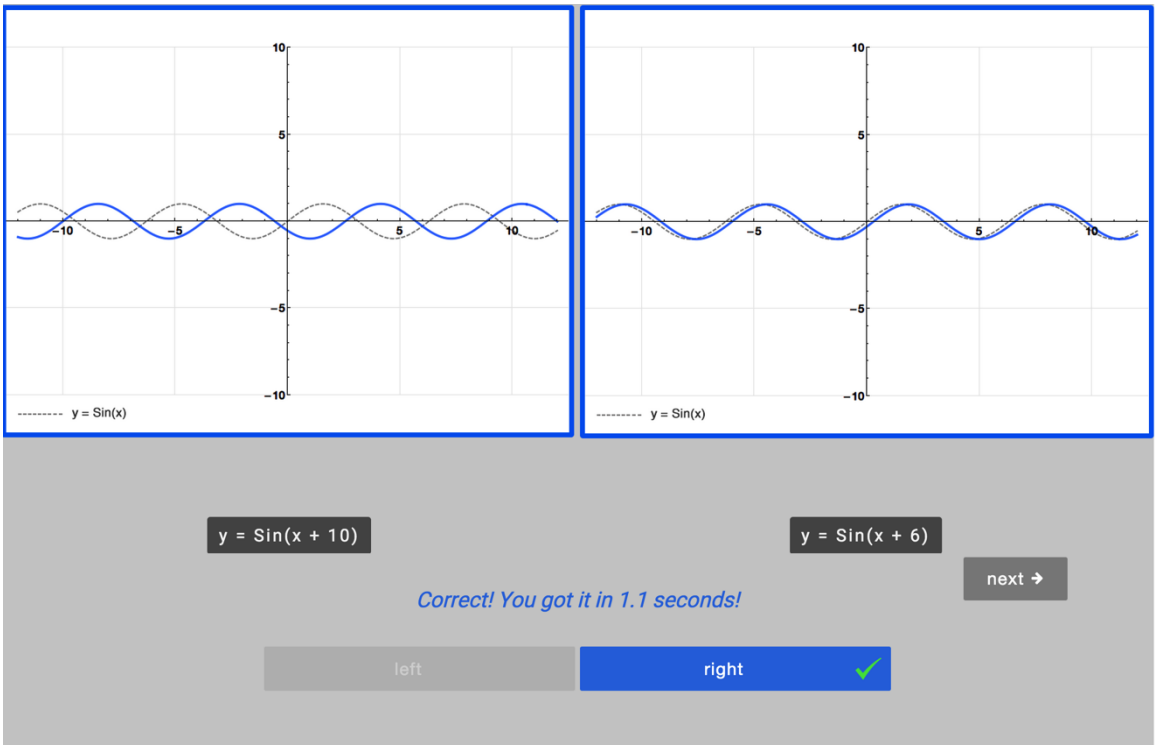


Figure 6.2. Feedback of a sample AA trial

Survey

In this survey, for exploratory purposes we added 4 multiple-choice questions, each asking participants to describe the transformation seen in a particular equation. *Appendix A* contains these questions. Perceptual learning does not develop separately from declarative knowledge (Kellman & Massey, 2013), and one possibility is comparison practice may highlight similarities and differences of transformations in a way that facilitates the development of declarative knowledge.

Procedure

The procedure was identical to that used in Experiment 4[14]. In the *AL/AC* and *AL/NC* condition, each participant's pattern of errors triggered either an AB or an AA comparison trial. In the *AL/AC* condition, the type of comparison triggered was based on errors made at the subcategory level. When participants twice in a row chose instances of one consistent wrong category, they were given an AB trial. For example, when the target subcategory was *y*-shifting-upward, and a participant picked twice (out of 3 times) *y*-shifting downward (e.g., $y = exp(x) + 2$) in one trial and $y = exp(x) + 4$ in another), they received an AB trial showing a *y*-shifting upward graph and a *y*-shifting downward graph (e.g., $y = exp(x) - 2$) and $y = exp(x) + 3$, respectively). The instances from each subcategory were selected at random. Participants were asked to identify, by indicating "left" or "right", the function that matches a given equation (e.g., "$y = exp(x) - 2$)").

In the same training condition, when participants incorrectly chose two instances from two different subcategories for the target subcategory, they saw an AA trial. For example, when the target category was *x*-scaling with expansion, and participants in one trial picked an *x*-

[14] Experiment 4 and 6 were conducted at the exact same time. The *single* condition from Experiment 5 is presented here as the *AL* condition in Experiment 6.

shifting function (e.g., $y = sin(x + 2)$) and in another picked a *y*-shifting function (e.g., $y = 2 + sin(x)$), they were shown two instances of *x*-scaling with expansion (e.g., $y = sin(2x)$ and $y = sin(3x)$). They were asked to choose among the plotted function the one that matches a given equation (e.g., "$y = sin(2x)$"). Accuracy feedback was provided after each of these comparison trials, and the graphs were replaced with their respective contrastive version. The PALM kept track of the past 3 instances of each category, so patterns with no or one intervening correct response could have triggered a comparison trial. The performance on these trials did not contribute toward the learning criteria. Participants returned to an AL trial immediately after each comparison trial.

The *AL/NC* group received the same adaptive learning paradigm with adaptive classification trials, and the comparison trials were also triggered by a pattern of error. Crucially, however, the type of comparison trials and the subcategories presented for comparison were selected at random. Thus, it was possible that after a classification trial on *y*-shifting-upward, participants could have received an AB trial showing *y*-scaling-expansion and *x*-scaling-expansion. The total number of comparison trials and the ratio of AB and AA trials (33:67) were matched between the *AL/AC* and *AL/NC* conditions. The ratio of AB and AA trials was determined from pilot data.

Overview of Analyses and Expected Results

Similar with Experiment 4, we only collected and analyzed data from participants who have completed all phases of the study (N = 24 per condition), all of whom did not experience technical difficulties and did not self-report to have looked up the materials at any point during the study. There were 16 others in the *AL* condition, 23 in the *AL/AC* condition, and 13 in the *AL/NC* condition who started the PALM but dropped out before reaching learning criteria.

Based on prior studies, we expected strong improvements from pretest to immediate posttest and high maintenance of learning at delayed test. Similar to Experiment 5, we considered both AL and comparison trials toward the calculation of trial efficiency scores. To determine whether comparison trials reduce the number of AL trials needed for achieving learning criteria, we also compared efficiency with just the AL trials completed. We expected the comparison trials, particularly adaptive comparisons, to support training efficiency, by enhancing transfer accuracy at posttests and/or by lowering the total number of trials needed achieve learning criteria.

RESULTS

Efficiency

Efficiency by Trials

Figure 6.3a displays the efficiency by total trial for each condition. A 2 phase (pre-post, pre-delayed) x 3 condition (*AL/AC, AL/NC, AL*) ANCOVA with pretest accuracy as the covariate showed neither significant main effects nor interactions, F's < 2, p's > .10. Overall, the *AL/AC* condition led to marginally greater trial efficiency than *AL/NC* (.0015 vs. .0010, $t(46) = 1.86$, $p = .07$, $d = .54$. There was no other condition differences on overall trial efficiency (*AL/AC* vs. *AL, t(46) = 1.34, p = .19; AL/NC* vs. *AL*, *t(46) < 1, p > .20).*

Other planned pairwise comparison showed that the only notable effect was that the *AL/AC* produced higher efficiency than *AL/NC* at delayed test with a medium effect size (.0013 vs. .0008, respectively), $t(46) = 2.02$, $p = .049$, $d = .58$. There were no other condition differences on pre-post and pre-delayed efficiency, $t(46) < 1.5$, p's > .15.

We also examined the efficiency with just *AL* trials (accuracy gain divided by total number of AL trials), which confirmed a main effect of condition, $F(2,69) = 4.35$, $p = .02$, $\eta^2_p =$

.11. The *AL/AC* had reliably higher AL efficiency than the *AL* condition overall, $t(46) = 2.73$, $p = .009$, $d = .64$, suggesting that the comparison trials were effective for reducing the number of AL trials needed. This held with large and medium effect sizes for pre-post AL efficiency, $t(46) = 2.73$, $p = .009$, $d = .80$, and pre-delayed AL efficiency, $t(46) = 2.40$, $p = .02$, $d = .62$, respectively. Not just any kind of comparisons was helpful at the delay, however. *AL/NC* did marginally better than *AL* on pre-post AL efficiency, $t(46) = 1.74$, $p = .09$, $d = .54$, but there was no difference between these two conditions on pre-delayed AL efficiency, $t(46) < 1$, $p > .20$, nor on overall AL efficiency, $t(46) = 1.36$, $p = .18$. Furthermore, the *AL/AC* condition did not do better than the *AL/NC* on pre-post AL efficiency, $t(46) = 1.29$, $p > .20$, but it was marginally better on pre-delayed, $t(46) = 1.81$, $p = .08$, $d = .53$, and overall efficiency, $t(46) = 1.67$, $p = .10$, $d = .52$.

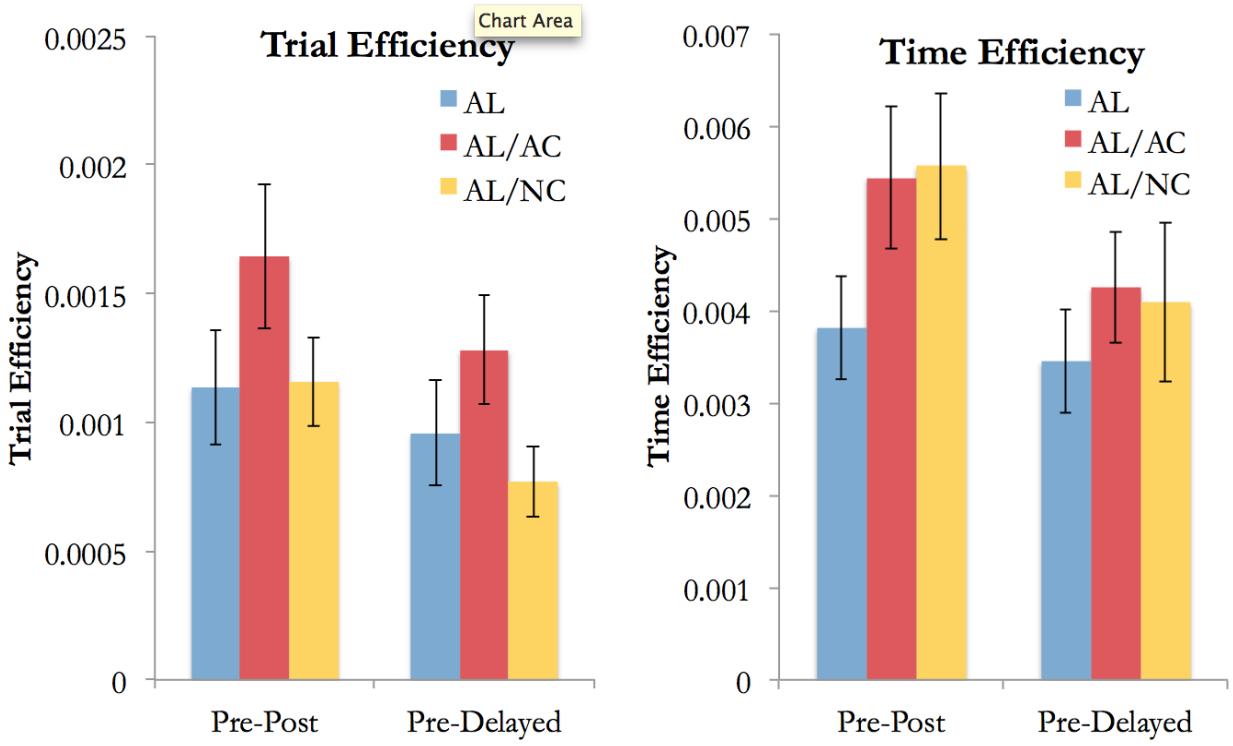


Figure 6.3. Efficiency (a) by total trial and (b) by time.

By Assessment Item Types

The only assessment type for which there were reliable condition differences was novel items of trained function (TF/NI) items. The 2 phase (pre-post, pre-delayed) x 3 condition (*AL/AC, AL/NC, AL*) ANCOVA with pretest TF/NI accuracy as the covariate confirmed a marginally significant main effect of condition, $F(2,68) = 2.45$, $p = .09$, $\eta^2_p = .07$. The *AL/AC* condition had marginally overall higher trial efficiency than the *AL/NC* condition (.0014 vs. .0009, $t(46) = 1.83$, $p = .07$, $d = .53$). This difference was significant at delayed test (.0028 vs. .0016, $t(46) = 2.46$, $p = .02$, $d = .68$) and marginal at immediate posttest (.0034 vs. .0022, $t(46) = 1.85$, $p = .07$, $d = .55$). *AL/AC* did not differ reliably from *AL,* $t(46) < 1$, p's > .20, and *AL/NC* did not differ reliably from *AL,* $t(46) < 1.4$, p's > .18. *Appendix J* provides more details of these results.

Efficiency by Time

Figure 6.3b displays the time efficiency by condition. The only condition differences in time efficiencies occurred at immediate posttest. At immediate posttest, both *AL/AC* and *AL/NC* had marginally higher time efficiency than *AL*, $t(46) = 1.71$, $p = .09$, $d = .53$, and $t(46) = 1.81$, $p = .08$, $d = .49$, respectively. There were no differences in time efficiency at delayed test, $t(46) < 1$, p's > .20.

By Assessment Item Types

There were no notable condition differences in time efficiency for each item type, $t(46) < 1.4$, p's > .17. However, and interestingly, on combination function (CF) items, both the *AL/AC* and *AL/NC* had higher overall CF time efficiency than *AL* with medium and large effect sizes *(*.0045 and .0039 vs. -.0008, $t(46) = 2.32$, $p = .03$, $d = .67$, and $t(46) = 2.83$, $p = .007$, $d = .82$,

respectively. At pre-post, the effect of *AL/AC* over *AL* was marginal (.0026 vs. -.0019), $t(46) = 2.01$, $p = .05$, $d = .58$, and the effect of *AL/NC* was significant (.0035 vs. -.0019), $t(46) = 2.49$, $p = .02$, $d = .72$. At pre-delayed, both *AL/AC* and *AL/NC* were reliably greater than *AL* with medium effect sizes (.0063 and .0043 vs. .0003), $t(46) = 2.26$, $p = .03$, $d = .65$, and $t(46) = 2.50$, $p = .02$, $d = .72$, respectively. The differences between *AL/AC* and *AL/NC* were not statistically significant, $t(46) < 1$, $p > .20$. *Figure 6.4* shows the average CF time efficiency by condition.

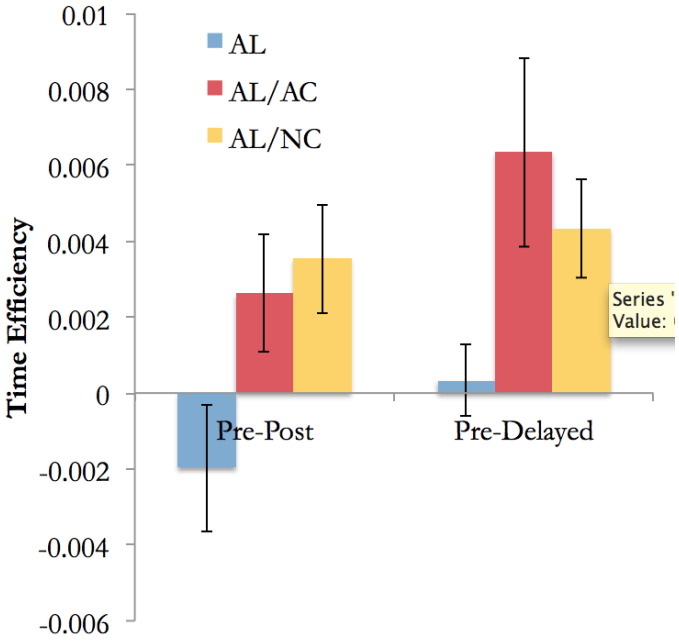


Figure 6.4. Mean time efficiency by condition for combination function (CF) items.

Accuracy

All Items

Figure 6.5a shows the mean accuracy on all items. A 3 phase (pre, post, delayed) x 3 condition (*AL/AC, AL/NC, AL*) ANOVA confirmed that participants from all three conditions experienced large improvements that persisted a week later even on items that were never shown in the training with a main effect of phase, $F(2,138) = 150.73$, $p < .001$, $\eta^2_p = .69$. The improvement from pretest to immediate posttest and to delayed test had very large effect sizes,

$t(71) = 16.29$, $p < .001$, $d = 2.10$, and $t(71) = 16.29$, $p < .001$, $d = 1.81$, respectively. There was some forgetting between the immediate and delayed test with a small effect size, $t(71) = 3.28$, $p < .01$, $d = .38$.

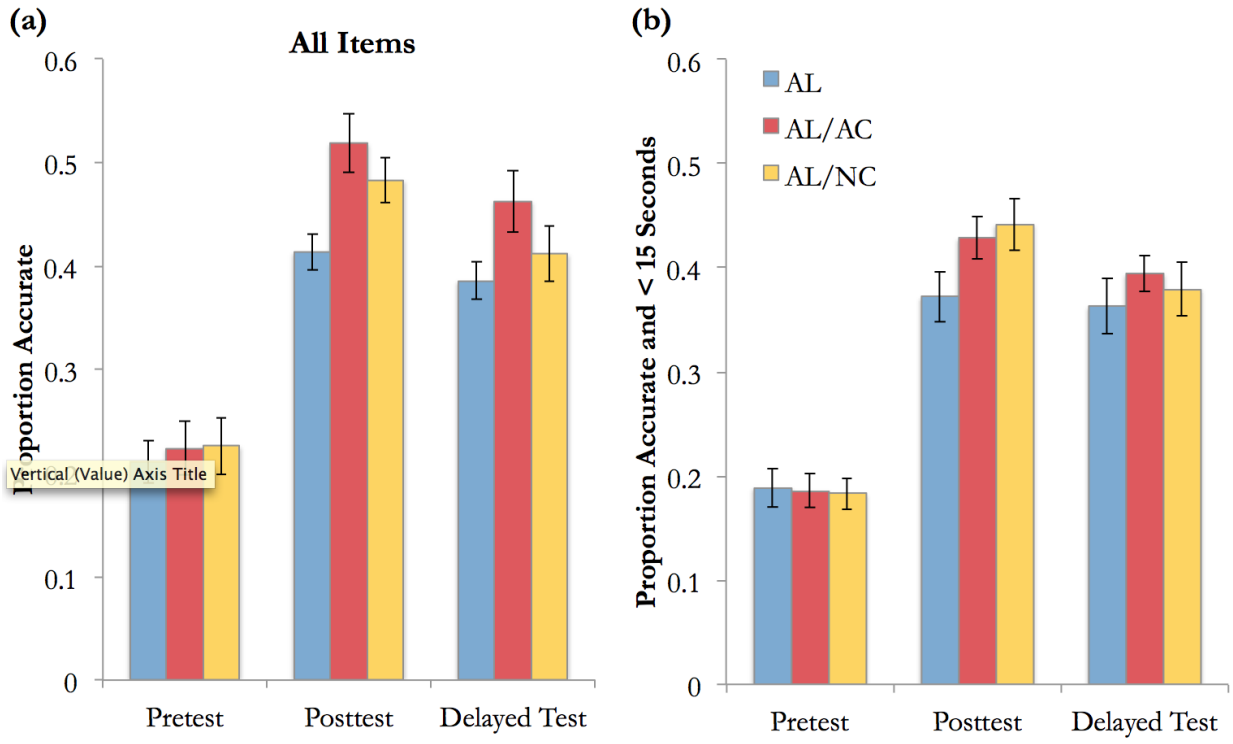


Figure 6.5. Mean (a) accuracy and (b) fluent accuracy on all items.

As we expected, there was also a marginal main effect of condition, $F(2,69) = 2.92$, $p = .06$, $\eta^2_p = .08$. The *AL/AC* group had greater overall accuracy ($M = .40$, $SD = .09$) than the *AL* group ($M = .34$, $SD = .10$), $t(46) = 2.36$, $p < .05$, $d = .68$, and there were no reliable differences between *AL/AC* and *AL/NC* ($M = .37$, $SD = .09$), $t(46) = 1.07$, $p > .20$, nor between *AL/NC* and *AL*, $t(46) = 1.36$, $p = .18$. There was no phase x condition interaction, $F(4,138) = 1.77$, $p = .14$.

Planned comparisons showed that there were no reliable condition differences at pretest, $t(46) < 1$, $p > .20$. At immediate posttest, *AL/AC* did significantly better than *AL* (.52 vs. .41), $t(46) = 2.72$, $p = .009$, $d = .78$). *AL/NC* did marginally better than *AL* (.48 vs. .41), $t(46) = 1.74$, $p = .09$, $d = .50$. At delayed test, *AL/AC* also outperformed *AL* (.46 vs. 39), $t(46) = 2.19$, $p = .03$, d

= .63, but there was no reliable difference between *AL/NC* and *AL* (.41 vs. .39), $t(46) < 1$, $p > .20$. *AL/AC* did not differ reliably from *AL/NC* on either immediate posttest, $t(46) < 1$, $p > .20$ or delayed test, $t(46) = 1.46$, $p = .15$.

Accuracy gain

Accuracy gains showed the same pattern. Adaptive comparisons led to greater learning gain in the *AL/AC* condition than the *AL* condition with a medium effect size (.27 vs. .19), $t(46) = 2.57$, $p = .01$, $d = .74$. This difference was statistically reliable at immediate pre-post gain (.29 vs. .20), $t(46) = 2.51$, $p = .02$, $d = .72$, and was marginally reliable at pre-delayed gain (.24 vs. .18), $t(46) = 1.98$, $p = .05$, $d = .57$. There were no differences between *AL/AC* and *AL/NC,* $t(46) = 1.44$, $p = .16$, and between *AL/NC* and *AL*, $t(46) < 1$, $p > .20$.

By Assessment Item Types

Figure 6.6a shows the mean accuracy on trained items (TI). *AL/AC* led to higher overall TI accuracy than *AL* (.43 vs. .35), $t(46) = 2.61$, $p = .01$, $d = .75$. This difference held at immediate posttest (.63 vs. .48), $t(46) = 2.61$, $p = .01$, $d = .75$, but was not reliable at delayed test, $t(46) = 1.33$, $p = .19$. In effect, *AL/AC* had marginally higher pre-post TI accuracy gain than *AL,* $t(46) = 1.70$, $p = .096$, but not higher pre-delayed TI accuracy gain, $t(46) < 1$, $p > .20$.

AL/NC also led to marginally greater overall TI accuracy than *AL* (.41 vs. 35), and $t(46) = 2.02$, $p = .05$, $d = .58$, but this may be partly driven by *AL/NC* having marginally higher pretest TI accuracy than *AL (*.22 vs. .16), $t(46) = 1.79$, $p = .08$, $d = .52$. The difference did not hold reliably at either posttests, $t(46) = 1.64$, $p = .11$, or delayed test, $t(46) < 1$, $p > .20$. As a result, in terms of TI accuracy gain, there were no differences between *AL/NC* and *AL*, $t(46) < 1$, $p < .20$. There were no differences between *AL/AC* and *AL/NC* on TI and TI gain, $t(46) < 1.1$, $p > .20$.

This pattern was found for both Exponential TI and Sine TI, except that the difference between *AL/NC* and *AL* on overall Exponential TI was not statistically reliable, $t(46) = 1.32$, $p = .19$.

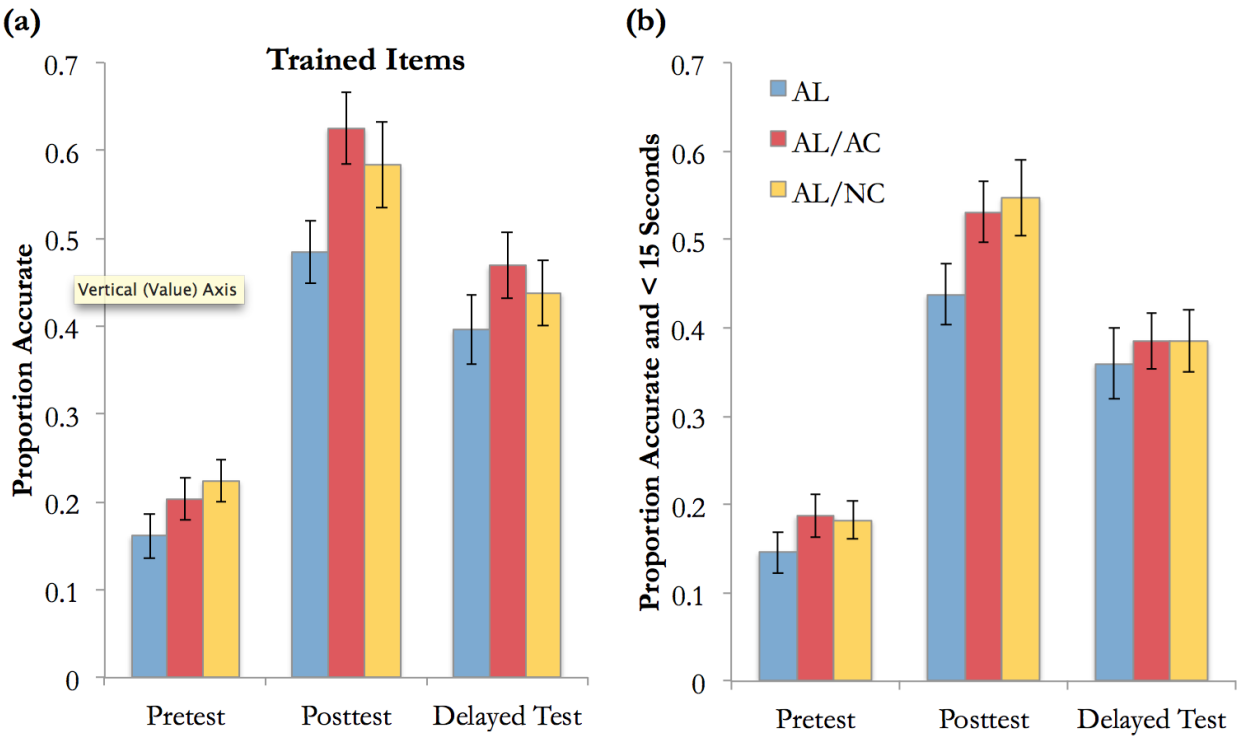


Figure 6.6. Mean (a) accuracy and (b) fluent accuracy on trained items.

On trained functions, novel items (TF/NI), transfer ability was particularly strong for those in the *AL/AC* group, $F(2,69) = 5.79$, $p = .005$, $\eta^2_p = .14$. They performed better overall than both the *AL*, $t(46) = 3.47$, $p = .001$, $d = 1.00$, and the *AL/NC* groups, $t(46) = 2.29$, $p = .03$, $d = .66$. In terms of TF/NI accuracy gain, *AL/AC* did marginally better than *AL* on pre-post, $t(46) = 1.80$, $p = .08$, $d = .52$, but not better on pre-delayed gain or on overall gain, $t(46) < 1$, p's $> .20$. *AL/AC* did not have higher pre-post nor pre-delayed accuracy gain than *AL/NC, t(46) < 1.6, p*'s $> .13$, but they had marginally higher overall accuracy gain, $t(46) = 1.75$, $p = .09$, $d = .51$. There was no reliable difference between *AL/NC* and *AL*, $t(46) = 1.02$, $p > .20$. *Figure 6.7a* shows the average TF/NI by condition. This pattern was similar for Exponential and Sine TF/NI, except

that *AL/AC* did not differ from *AL/NC* on Sine TF/NI, $t(46) < 1$, $p > .20$. *Appendix J* contains more details of these analyses.

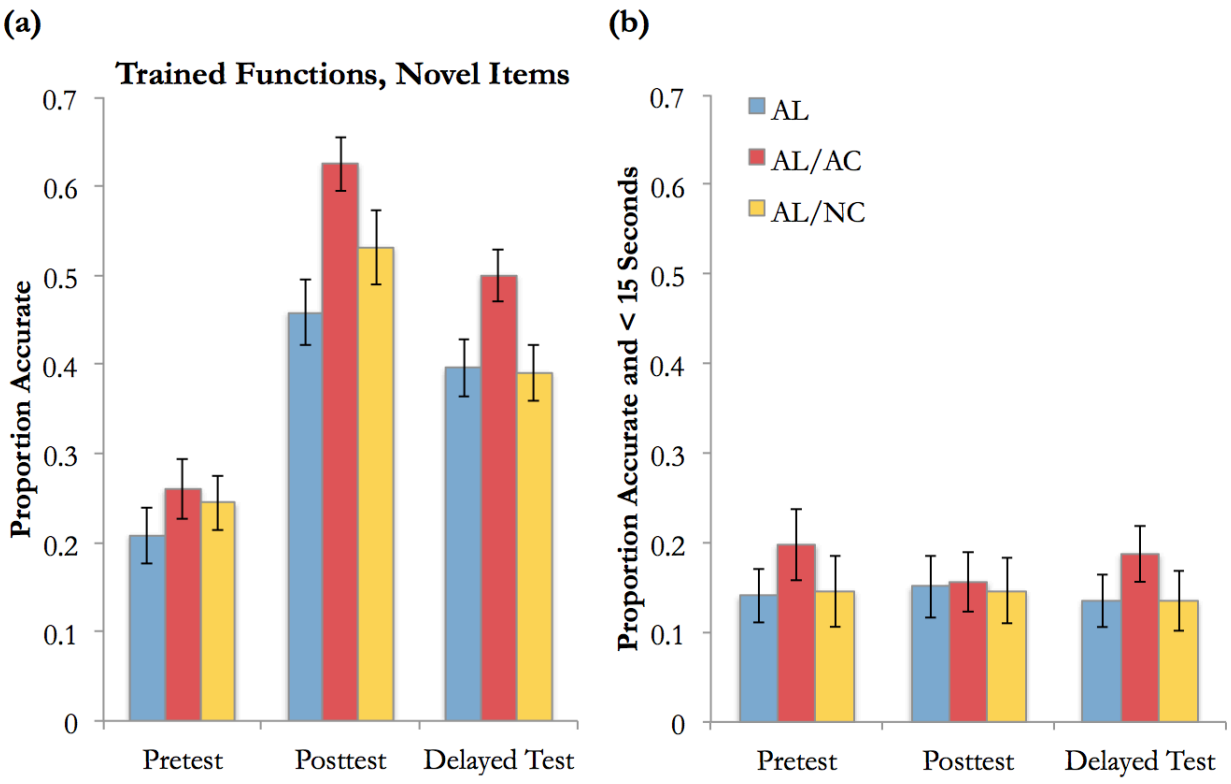


Figure 6.7. Mean (a) accuracy and (b) fluent accuracy on Trained Functions, Novel Items.

There were no condition differences on untrained function (UF) items, $t(46) < 1.4$, p's $> .18$, except that *AL/AC* had marginally higher overall UF accuracy than *AL/NC* (.32 vs. .26), $t(46) = 1.66$, $p = .10$, $d = .48$. Interestingly, however, there were condition differences on Cosine items but not on Logarithmic items. *Figure 6.8a* shows the mean accuracy on Cosine items, and *Figure 6.8b* shows the mean accuracy on Logarithm items. On Cosine items, the *AL/AC* condition did better than the *AL* condition, 43% vs. 31%, $t(46) = 2.21$, $p < .05$, $d = .64$, and than the *AL/NC* condition, 43% vs. 34%, $t(46) = 1.74$, $p = .09$, $d = .50$, both with medium effect sizes. There was no reliable difference between the *AL* and the *AL/NC* condition, $t(46) < 1$, $p > .20$.

There were no condition differences on Logarithmic items, $t(46) < 1.2$, $p > .20$, but interestingly, participants in the *AL/AC* and *AL/NC* groups showed reliable learning gain from pretest to delayed test on Logarithmic items, $t(46) = 2.22$, $p = .04$, $d = .45$, and $t(46) = 2.90$, $p = .01$, $d = .59$, respectively. This was not found with the *AL* group, $t(46) = 1.52$, $p = .14$. More details of these analyses are in *Appendix J*.

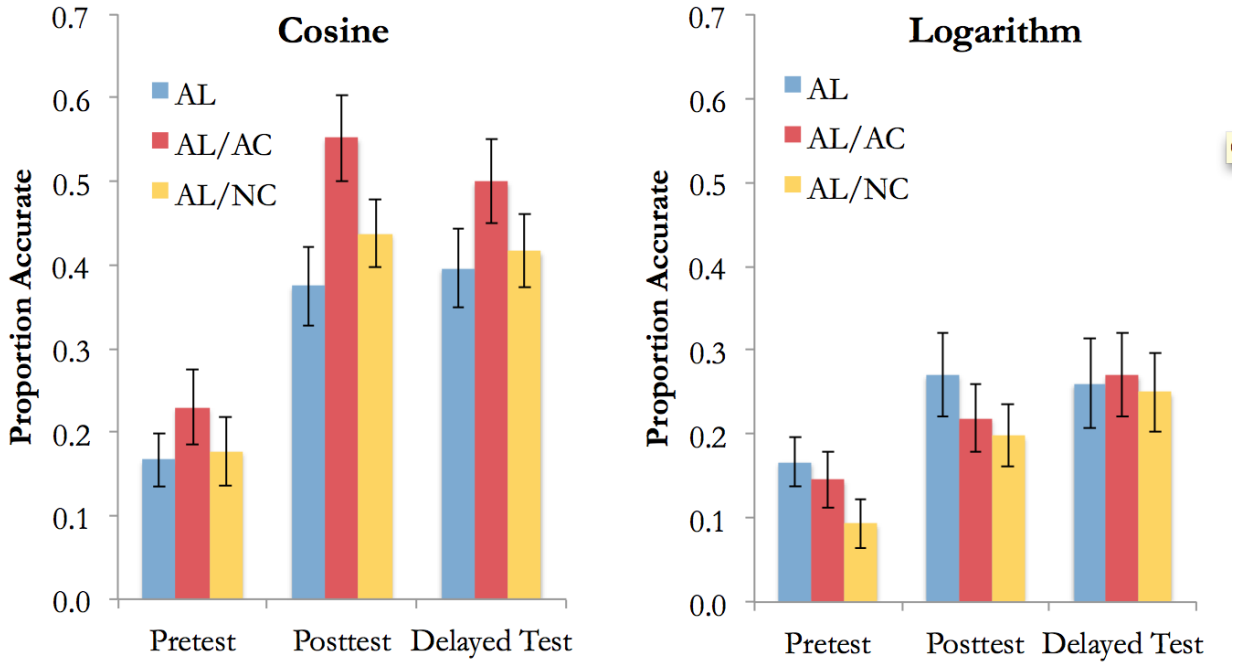


Figure 6.8. Mean accuracy on (a) Cosine and (b) Logarithmic UF items.

On combination function (CF) items, the *AL/NC* marginally outperformed *AL/AC* on overall accuracy, $t(46) = 2.01$, $p = .05$, $d = .60$. This is the only instance when *AL/NC* did better than *AL/AC*. However, this may partly be due to *AL/NC* having numerically (but not reliably) higher pretest CF at the start than *AL/AC*, $t(46) = 1.64$, $p = .11$, and higher immediate posttest, $t(46) = 2.08$, $p = .04$, $d = .67$, but there were no condition differences at delay test, nor on any accuracy gain measures, $t(46) < 1$, $p > .20$. But both types of training with comparisons led to greater accuracy gain on combination functions than the *AL* training, with marginal significance

but medium effect sizes (*AL/AC* vs. *AL*, $t(46) = 1.87$, $p = .07$, $d = .54$, *AL/NC* vs. *AL*, $t(46) = 1.73$, $p = .09$, $d = .50$). *Figure 6.8* shows the CF accuracy by condition.

Interestingly, the comparison conditions *AL/AC* and *AL/NC* produced significant learning gains from pretest to delayed test, $t(46) = 3.92$, $p < .001$, $d = .80$, and $t(46) = 2.84$, $p = .01$, $d =$, but the *AL* condition did not, $t(46) = 1.27$, $p > .20$.

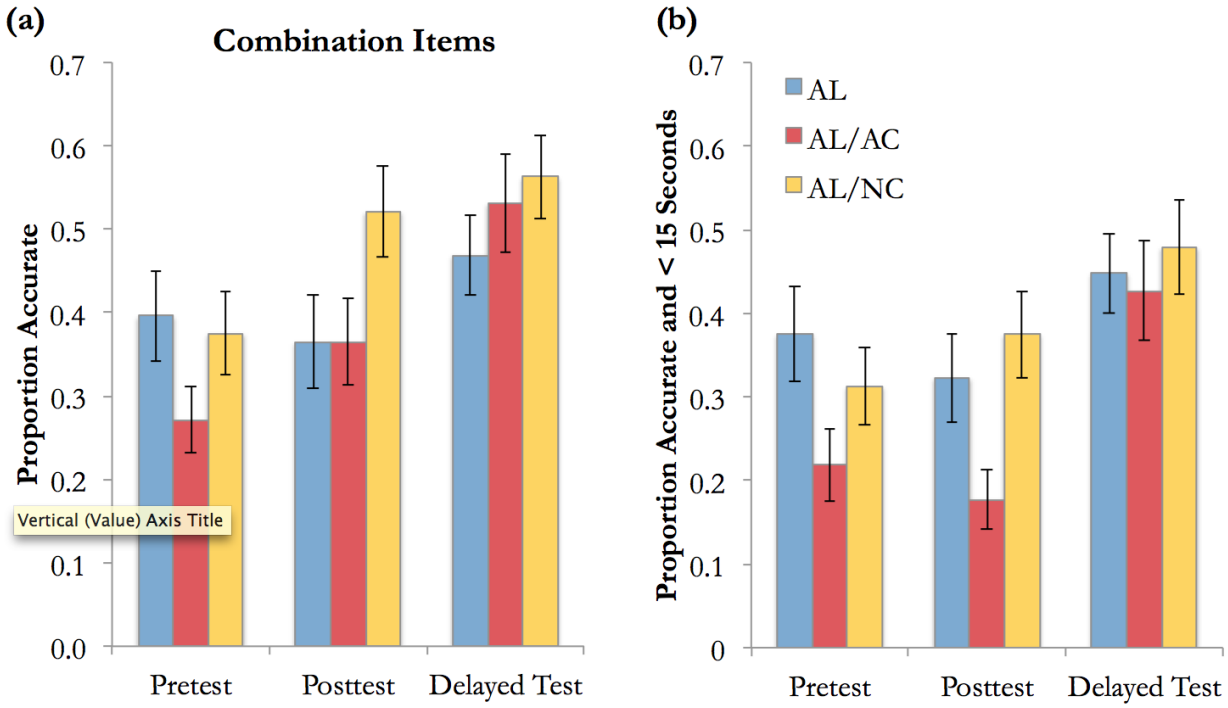


Figure 6.9. Mean (a) accuracy and (b) fluent accuracy on combination function items.

Fluency

Figure 6.5b shows the mean fluent accuracy on all items. Fluent accuracy showed similar patterns of condition differences as accuracy. The *AL/AC* condition led to greater performance than the *AL* condition on overall fluent accuracy (.40 vs. .34), $t(46) = 2.37$, $p = .02$, $d = .69$, and also on fluent accuracy gain (.27 vs. .19), $t(46) = 2.52$, $p = .02$, $d = .73$. The *AL/AC* and *AL/NC* did not differ on overall fluent accuracy (.40 vs. .37), $t(46) = 1.36$, $p = .18$, nor on overall fluent

accuracy gain (.27 vs. .22), $t(46) < 1$, $p > .20$. Similarly, *AL/NC* did not differ from *AL* on overall fluent accuracy, $t(46) = 1.08$, $p > .20$, nor fluent accuracy gain, $t(46) = 1.42$, $p = .16$.

On trained items (TI; *Figure 6.6b*), both *AL/AC* and *AL/NC* had marginally higher overall fluent accuracy than *AL* with medium effect sizes (.37 and .37 vs. .31), $t(46) = 1.91$, $p = .06$, $d = .55$, and $t(46) = 1.87$, $p = .07$, $d = .54$, respectively. On novel items of trained functions (TF/NI; *Figure 6.7b*), *AL/AC* had higher overall TF/NI fluent accuracy than *AL* with a large effect size (.40 vs. .32), $t(46) = 2.84$, $p = .007$, $d = .82$. *AL/NC* only did better than *AL* at immediate posttest, $t(46) = 2.19$, $p = .03$, $d = .64$, and not at delayed test, $t(46) < 1$, $p > .20$. There were no notable differences among conditions on untrained functions (UF), $t(46) < 1$, p's $> .20$. Lastly, on combination functions (CF; *Figure 6.9b*), both *AL/NC* and *AL* had overall higher fluent accuracy than *AL/AC* (39% and .38% vs. 27%), $t(46) = 2.31$, $p = .03$, $d = .67$, and $t(46) = 2.21$, $p = .03$, $d = .64$, respectively. This was likely because the *AL* and *AL/NC* groups started out with higher fluent accuracy on these items than the *AL/AC* group (38% and 31% vs. 22%, respectively, $t(46) = 2.20$, $p = .03$, $d = .63$ and $t(46) = 1.49$, $p > .10$). There was no difference between *AL/NC* and *AL*, and no condition differences on CF fluent accuracy gain, $t(46) < 1$, p's $> .20$, . There were no other reliable condition differences on these item types, p's $> .10$. *Appendix J* contains more details of these analyses.

Progression of Learning

Table 6.1 contains the training means by condition. Conditions did not differ in the total number of trials nor in the amount of time needed for reaching learning criteria, p's $> .10$. Although adaptive comparison practice did not reduce the total number of training trials, it reduced the number of AL trials needed to reach learning criteria. There was a significant

difference in the number of AL trials to complete the module between the *AL/AC* and the *AL* condition (185.83 vs. 260.33 trials, respectively), $t(46) = 2.12$, $p = .03$, $d = .64$. There were no reliable differences between the other conditions on the number *AL* trials needed for completion, p's > .10.

Condition	Minutes on Training	AL trials	AA trials	AB trials	AL accuracy	AA accuracy	AB accuracy
AL	40.67	154.5			.53		
	(2.72)	(8.73)	--	--	(.02)		
AL/AC	42.50	119.5	25.71	11.08	.58	.71	.52
	(2.80)	(6.86)	(3.24)	(1.31)	(.02)	(.03)	(.04)
AL/NC	47.33	131.4	27.79	13.58	.57	.72	.62
	(2.91)	(7.16)	(3.22)	(1.96)	(.02)	(.03)	(.05)

Table 6.1. Training means by condition. Standard errors are in parentheses.

Overall there were condition differences on accuracy of AL trials, $F(2,69) = 6.19$, $p = .003$. The *AL/AC* group had overall higher accuracy on AL trials than both *AL* and *AL/NC* groups with large and medium effect sizes (42% vs. 32% and 36%, respectively, $t(46) = 3.14$, $p = .003$, $d = .91$, and $t(46) = 2.07$, $p = .045$, $d = .60$, respectively). These differences, however, may be due to a speed-accuracy trade-off. The *AL/AC* group took longer to reach the correct answers on AL trials than both the *AL* and *AL/NC* groups with medium effect sizes (8.33 seconds vs. 6.39 seconds and 6.46 seconds, $t(46) = 2.44$, $p = .02$, $d = .60$, and $t(46) = 2.52$, $p = .02$, $d = .73$, respectively).

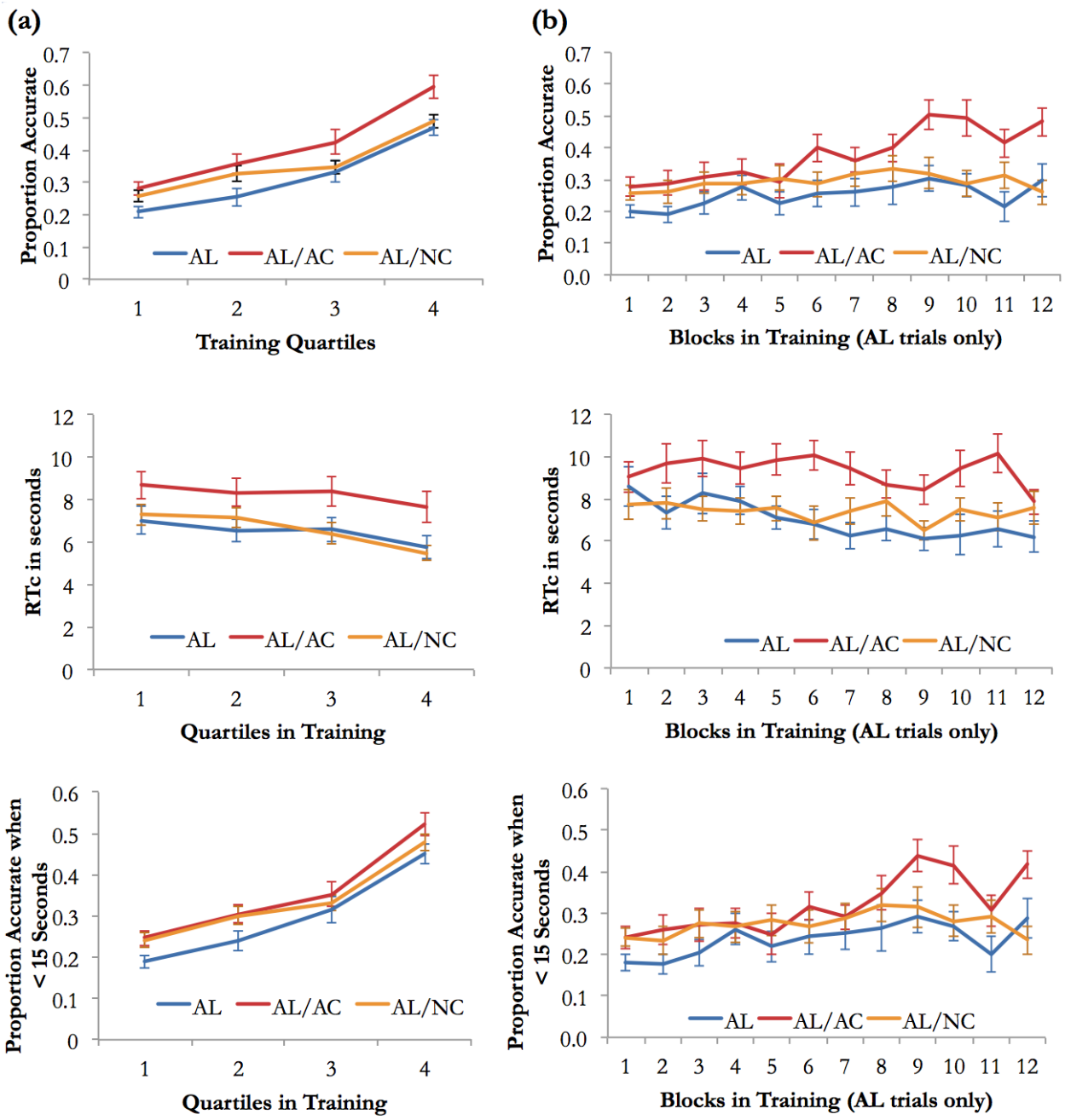


Figure 6.10. Mean accuracy, RTc, and fluent accuracy on AL trials (a) by quartiles and (b) by blocks in the training. Each block had 12 total trials (AL and comparison trials). Note that because only performance on AL trials is plotted, most blocks in *AL/AC* and *AL/NC* had fewer than 12 AL trials. Because everyone was trained toward learning criteria, not all participants completed 12 blocks (some had more, some had fewer).

Figure 6.10 shows the accuracy, RTc, and fluent accuracy by quartiles and by blocks on AL trials. *AL/AC* had higher accuracy than *AL* across all four quartiles in the training, and *AL/AC* also was generally slower than *AL* across all four quartiles in the training, with medium to large effect sizes, $t(46) > 1.84$, $p = .02$ to $.07$, $d = .53$ to $.81$. We also examined these differences in terms of blocks and found that at Block 1, *AL/AC* did not differ from *AL*, $p > .10$, but *AL/AC* already had higher RTc (10.67 seconds vs. 8.10 seconds, $t(46) = 1.98$, $p = .05$, $d = .57$.

The *AL/AC* group did not differ from *AL/NC* in accuracy during the first 2 quartiles, but they did better than the *AL/NC* group during the latter two, $t(46) = 1.84$, $p = .07$, $d = .53$, and $t(46) = 2.52$, $p = .02$, $d = .73$, for the 3rd and 4th quartiles, respectively. *AL/AC* also took longer than *AL/NC* to get each question correct on the 1st, 3rd, and 4th quartiles, $t(46) > 1.75$, $p = .01$ to $.09$, $d = .51$ to $.79$. There were no differences between these two conditions on fluent accuracy in any quartiles, p's $> .10$. Taken together, the lack of differences in accuracy and RTc earlier on in the training suggested that condition effects were due to the training. However, the *AL/AC* group generally spent longer on each trial, suggesting that some of the effects seen during the training may be due to individual differences in overall speed. *Appendix J* contains more details of these analyses.

When this progress of learning was evaluated in terms of fluent accuracy, both *AL/AC* and *AL/NC* had higher fluent accuracy than *AL* (36% and 34% vs. 30%, $t(46) = 2.39$, $p = .02$, $d = .69$ and $t(46) = 1.84$, $p = .07$, $d = .53$, respectively). There was no reliable difference between *AL/NC* and *AL/AC*, $p > .10$.

Survey Questions

Metacognitive Questions

There were no condition differences on questions in the survey, p's > .05, with one exception. Participants differed in their judgments of how much they would remember at delayed test, $F(2,69) = 4.35$, $p < .05$. Both groups trained with comparisons judged that they would remember more at delayed test than the *AL* condition. These differences were reliable (out of 6, *AL/AC* ($M = 3.54$, $SD = 1.35$) vs. *AL* ($M = 2.54$, $SD = 1.14$), $t(46) = 2.77$, $p < .01$, $d = .80$, and the difference between *AL/NC* and *AL* was marginally significant, *AL/NC* ($M = 3.17$, $SD = 1.05$) vs. *AL*, $t(46) = 1.98$, $p = .05$, $d = .57$.

Describe the Transformation Items

These 4 questions asked participants to select the correct description of 4 types of transformations used in the training given an equation and its canonical function (i.e., From $y = sin(x)$, how do we get $y = sin(x + 2)$?). *Appendix B.2* contains the survey questions. *Figure 6.11* shows the mean accuracy on these items. Interesting, most people did not do well on this task despite having reached learning criteria, confirming the intuitive, implicit aspects of perceptual learning. Notably, the two comparison conditions did indeed do better than the *AL* condition at delayed test. There were condition differences on the immediate posttest description accuracy, $F(2,69) = 3.16$, $p = .049$, and delayed test description accuracy, $F(2,69) = 3.39$, $p = .04$. *AL/AC* had higher accuracy than *AL* at immediate posttest (54% vs. 33%), $t(46) = 2.56$, $p = .01$, $d = .74$ and at delayed test (45% vs. 28%), $t(46) = 2.35$, $p = .02$, $d = .68$. *AL/NC* did not do reliably better than *AL* at immediate posttest, $t(46) = 1.62$, $p = .11$, but *AL/NC* was reliably better than *AL* at delayed test (42% vs. 28%), $t(46) = 2.07$, $p = .04$, $d = .60$. There were no reliable differences between the two comparison conditions at post and delayed tests, $t(46) < 1$, p's > .20.

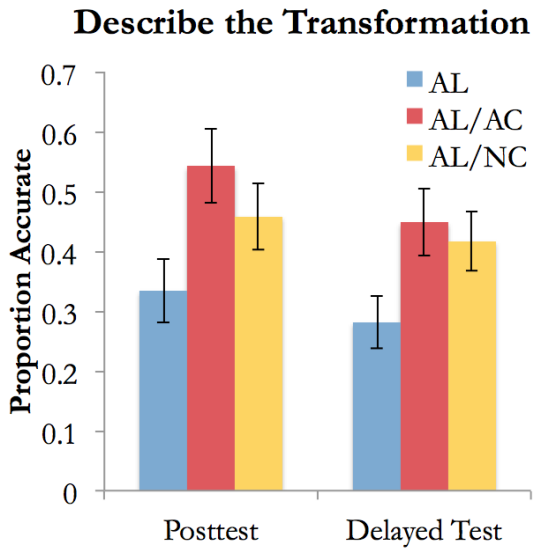


Figure 6.11. Accuracy on describe the transformation survey items.

DISCUSSION

As expected, and similar to prior studies, all PALM training conditions produced large and persistent learning gains, even on untrained items and untrained function families. In terms of our efficiency measure, the *AL/AC* condition outperformed *AL/NC* only at the delayed test. While the *AL/AC* training was not found to be more efficient than the *AL* training, it did reduce the total number of AL trials needed to reach learning criteria. There was some evidence that the *AL/AC* and *AL/NC* conditions were more efficient in terms of time than the *AL* condition, but this was a marginally significant effect found only at immediate posttest. Interestingly, however, both *AL/AC* and *AL/NC* were found to be reliably more time efficient than *AL* on combination function (CF) items.

In terms of *accuracy*, adaptive comparison trials greatly enhanced transfer accuracy over the no-comparison *AL* condition. *AL/AC* outperformed *AL* on all-items accuracy, particularly on trained items (TI), trained functions but novel items (TF/NI), Cosine untrained function (UF)

items, and even marginally so with combination function (CF) learning gains. The advantage of *AL/NC* over *AL* was limited; the only differences were marginally significant at TI, which were mainly driven by higher immediate posttest on Sine TI, and on CF learning gain.

It was notable that training with comparisons allowed participants to successfully transfer their learning to correctly classify combination functions, which were considerably more difficult than the items seen in the training. This was consistent with finding from Experiment 4, in which the *mixed* training with both *contrastive* comparisons and *single/AL* trials boosted participants' ability to transfer to combination function, confirming the benefit of comparison experiences.

The comparison trials in this experiment differed from those in Experiment 4 in important ways, and they led to stronger learning gains. In this experiment, comparisons were provided selectively. Rather than merely presenting the contrastive graphs and assuming that participants would engage in the comparison between the canonical function and the to-be-classified function, here we presented participants with two single graphs, showing either two of the type of transformation or of different types, and reserved the contrastive graphs for the feedback. Each comparison trial contained only two answer choices, and participants were asked to choose one of the graphs that matched a given equation. In doing so, we provided participants an opportunity to align the examples, but also directly (though implicitly) guided their attention to the similarities and differences between the two graphs to identify the features that were common (within-category comparison) or different (between-category comparison).

For example, seeing one graph showing $y = sin(x) + 4$ and another $y = sin(x) + 2$ side by side (in an AA trial), and being asked, "which one is $y = sin(x) + 4$?" may guide the learner to see that both functions cross the x-axis at the same point, but that one is higher up on the y-axis, at $y = 4$ versus $y = 2$. This should provide a major clue that $y = sin(x) + a$ involves an upward

shift on the *y*-axis by *a* units. Similarly, in an AB trial, one can imagine the same question but this time the second graph is $y = sin(x + 4)$. One may guess that one crossing the *y*-axis and another is identical, except that it is higher up on the *y*-axis at $y = 4$ should provide a significant clue that $\sin(x) + 2$ indicate an upward shift on the *y*-axis. In a between-category comparison, one can imagine the same question but this time the second graph could be $y = sin(x + 4)$ where the graph is shifted to the left by 4 units. Seeing that the two graphs are of equal size (without compression or expansion), but that one is shifted upward by 4 units, should lead the learner to knowing that one of the graphs involves a *y*-shifting transformation and the other involve some *x*-shifting. The trial feedback should be particularly informative in these cases, where having two contrastive graphs side-by-side can also highlight the differences between transformation types. These kinds of comparisons were notably different from that provided by a *contrastive* graph from Experiment 4, on which a canonical function was overlapped with the transformed function. In some cases, it may be more obvious which transformation was applied with contrastive graphs, such as those involving *y*-shifting. However, when it came to other transformation such as *x*-shifting, it was much less obvious what the transformation was and how much was the shift on the *x*-axis. As a result, training with comparisons in this experiment proved to be more effective than the baseline adaptive classification learning without any comparison trials.

These findings have educationally important implication regarding the effect of comparison. Although comparing multiple cases is considered as a high-quality instructional method (NCTM, 2000), comparison does not always guarantee better learning. To promote learning, what matters is not simply whether students compare examples but *how* they compare them, when, and what is being compared. Our research suggests that comparisons should be

designed to engage learners in the comparison process to extract the relevant structures for classifications, and that they address each individual's mistakes and confusions during the learning.

Furthermore, performance on the descriptive questions on the survey was generally low. This confirmed the notion that what is learned in perceptual learning is generally implicit. Interestingly, comparison practice enhanced learners' ability to verbally describe the learned transformations. One possibility is the comparison practice also engaged learners with declarative reasoning (e.g., "both of these were shifted upward") by spontaneously encourage description of the similarities and differences. This is consistent with prior evidence that people learn more when they were asked to explicitly compare (e.g., Gick & Holyoak, 1983; Gentner et al., 2003; Rittle-Johnson & Star, 2009). It is unlikely that the benefits of comparison are moderated by language (e.g. Hendrickson, Kachergis, Gureckis, Goldstone, 2010), but comparison practice may facilitate verbalization of the patterns. This is an exciting finding, suggesting that comparison practice may support both perceptual learning and declarative learning. Being able to classify patterns is important, but one should also be proficient in communicating their knowledge. Further research is needed to explore how comparisons may support explicit pattern recognition knowledge.

GENERAL DISCUSSION

Our findings confirmed that well-designed comparison practice can boost learners' extraction of the relevant patterns in the learning set, and allow them to flexibly transfer what they learn to more complex situations. In two very different domains, we found clear benefits of adaptive comparisons for pattern recognition learning and retention upon PALM completion. In both experiments, we found evidence supporting the superiority of *AL/AC* for learning and

transfer over *AL* alone, and that in terms of efficiency, it mattered that the comparisons were adaptive to learners' needs. The *AL/AC* condition was more efficient with both trials and time over the *AL/NC* condition in Experiment 5, and was more efficient in terms of trials at a delay over *AL/NC* in Experiment 6.

AL/AC was not always more effective than *AL/NC*, but there were some evidence suggesting that adaptive comparisons can produce higher learning gains on transfer items than non-adaptive comparisons. For example, the *AL/AC* had higher overall accuracy than *AL/NC* in Experiment 5, and in Experiment 6, *AL/AC* was more effective for producing TF/NI transfer (and marginally so for transfer to Cosine UF items). This benefit of adaptive comparison trials was noteworthy given that when participants completed the modules, they had already demonstrated mastery with the classification task. Yet, those who received adaptive comparison trials during practice seemed to have more flexible representations of the underlying structures for more effective transfer. It could be that because adaptive comparisons targeted the discriminations that learners have difficulty with, participants were able to learn the discriminations they need for mastery without wasting trials (or time) practicing with comparisons they did not have trouble with, which allowed them to develop greater fluency in processing those relations and perform better on transfer measures.

Between-category comparison may allow for picking out distinguishing features that separate the categories; within-category comparisons may support the pickup of the regularities that hold among members of each category. Prior studies have shown that the category structure and the level of feature variations determine the effectiveness of each type of paired-comparisons (e.g., Higgins & Ross, 2011; Ankowski, Vlach, and Sandhofer, 2012), such that within-category comparisons are more effective when category learning requires abstraction of the underlying

structure from many varying surface features, and that between-category comparisons are good for when the category structure is closely tied to the surface features. We combined the two types of comparisons in an adaptive paradigm, and tested with two very different domains involving different types of category structures; one arguably involves more variations in surface features than the other. We found that the benefit of adaptive comparisons held across these two different learning domains with different category structures.

This research has exciting theoretical and practical implications for understanding general learning mechanisms and for improving instruction in educational settings. Theoretically, these findings add to our understanding of the benefit of comparisons in learning. To the best of our knowledge, this was the first attempt to personalize the type of comparison to each learner's errors. Our paradigm proved effective two diverse domains, suggesting that we have tapped into a general learning mechanism that is important for enhancing learning, retention and efficiency of the training.

Much more research is needed to determine an optimal adaptive method. For example, participants only had one comparison trial each time it was triggered, and performance on those trials did not count toward learning criteria. Future studies may explore the potential additive effect of enabling a "mini" learning criterion on the comparison trials. For example, the comparison trials may be triggered the same way, but the learner must answer them correctly before exiting the comparison block and returning to the active classification trials. In this way, participants must show sufficient grasp of the similarities and differences among the problem instances before returning to the training.

Furthermore, this method is easily implemented from the point of view of an instructional designer. Other methods that use between- and within-category comparisons require the

instructor to know ahead of time the specific confusions or misunderstandings that students may have. The current adaptive comparison paradigm does this online. It is likely true that much of the time, students experience similar confusions, but this may not always be the case. This method, coupled with the existing adaptive learning paradigm, proved to be a potent combination that works across domains.

CONCLUSION

Not all comparisons reinforce perceptual learning, but adaptive comparisons paired with adaptive classification practice can produce durable learning gain and effective transfer, while doing so in an efficient way.

CHAPTER 7

Concluding Remarks

Summary

In six experiments, we examined several general learning principles at the nexus of perceptual learning and adaptive learning to accelerate aspects of learning that are difficult to address with traditional instruction. We did so in two different learning domains: ECG interpretation and mathematical transformations. We confirmed with a diverse sample of participants with different levels of interest and familiarity with the materials, that PALMs conferred genuine advances in learning while requiring modest training time (less than an hour for ECG and within a few hours for mathematics). While there were some nuances in the benefits of these variables for transfer and retention, when combining perceptual learning principles with adaptive learning technology, we found three robust principles:

1. That the combination of passive and active classification training was robustly more effective and efficient than training with only passive exposures
2. That training with only contrastive comparisons can limit learning and transfer
3. That the addition of adaptive comparisons to adaptive active classification training can enhance perceptual learning in an efficient way

Implications

This dissertation has exciting theoretical and practical implications for understanding general learning mechanisms and for improving instruction in educational settings. Our work builds upon this powerful yet natural ability of the human perceptual system to grow in its ability to isolate relevant detail, suppress irrelevancy, and pick up progressively deeper structure, as a result of appropriate kinds of learning experiences. This ability to see patterns can grow into

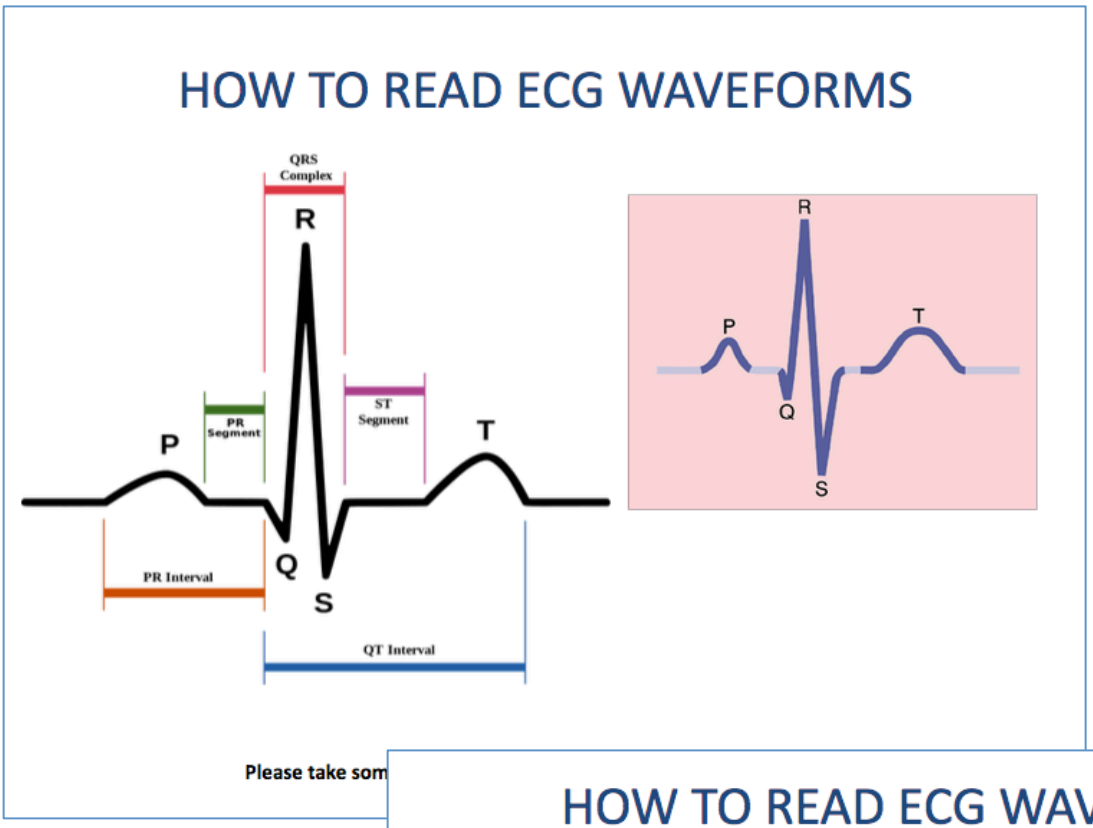
astonishing levels of sophistication, and is the cornerstone of advanced performance in many domains, such as science and mathematics, chess, aviation, and medicine. While our knowledge of the role of perceptual learning advances, we must account for and nurture students' natural tendency for perceptual learning. These experiments have focused on accelerating perceptual learning and enhancing adaptive learning to support transfer of learning to novel situations, to ensure that training has stable and lasting effect, and to do so in a more efficient way than ever before.

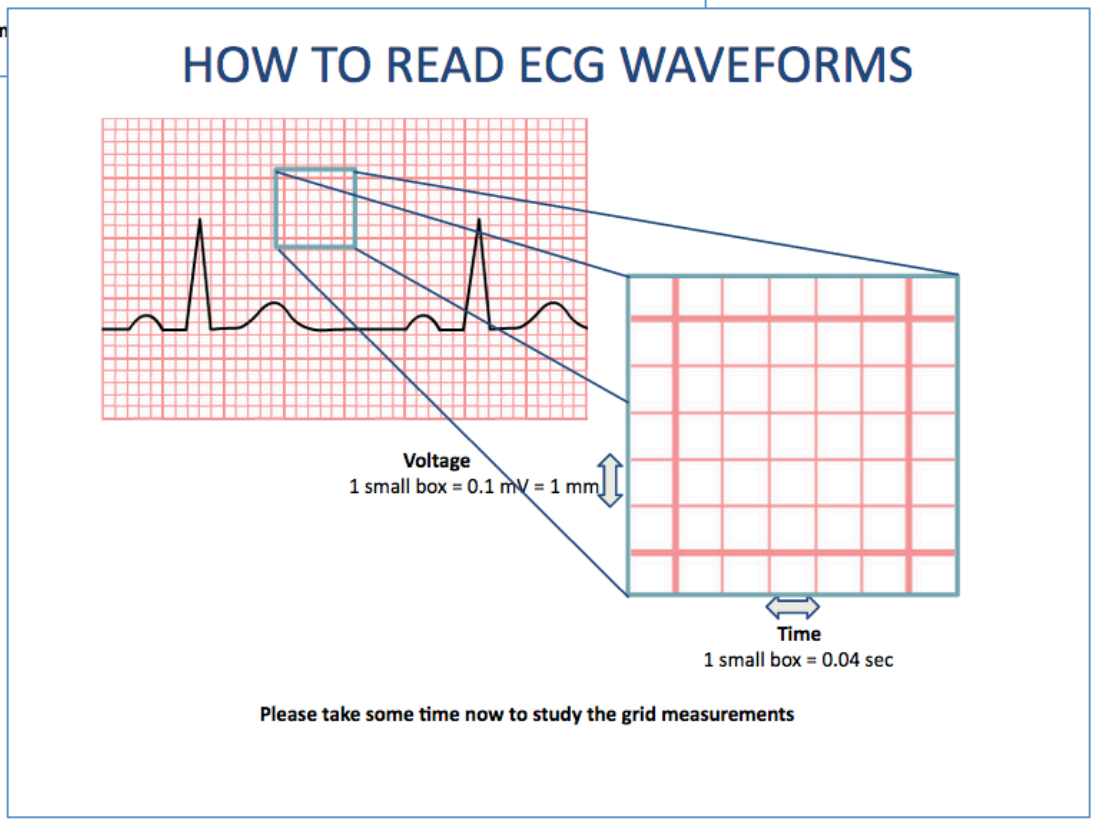
Theoretically, findings from this dissertation advance our understanding of the synergy of passive and active presentations, the benefit of comparisons in learning, and how to best personalize comparisons for each individual. By examining each of these effects in two separate domains, we have began to uncover general learning principles that bring about effective and efficient perceptual learning across domains.

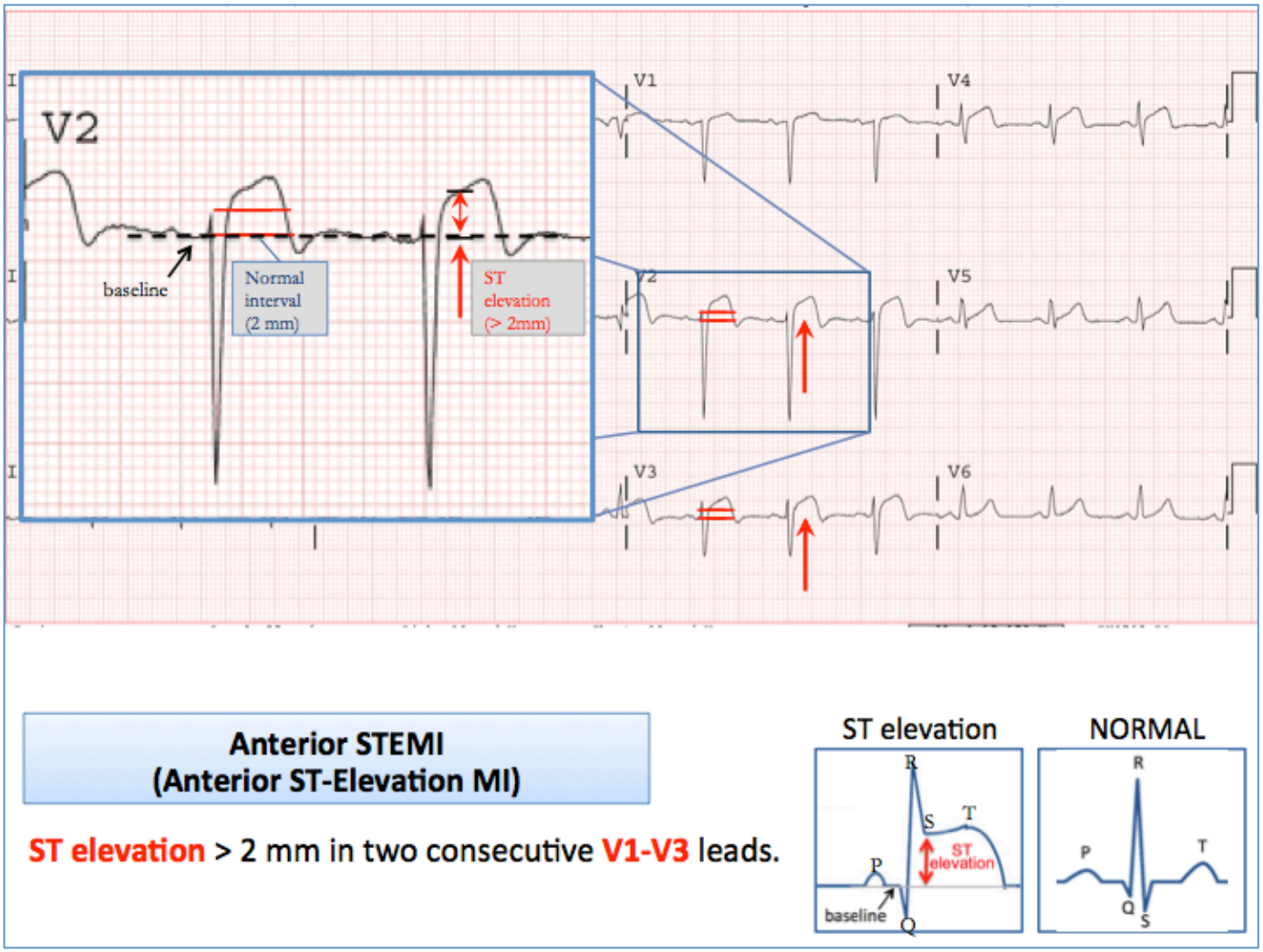
Practically, our findings can inform instructional designs (in terms of task formats, feedback displays, and personalized instruction) that shape students' perceptual processes so they can more efficiently and effectively process patterns, to apply their knowledge of facts and procedures to new situations. More research is needed to optimize the design and impact of PALMs, and to determine how best to integrate PALMs with traditional learning formats. While we perhaps can't *make* experts, we can provide optimal learning conditions and opportunities for students to achieve expertise with less time and effort than ever before.

APPENDIX A

A.1. Sample Primer Slides







Anterior STEMI
(Anterior ST-Elevation MI)

ST elevation > 2 mm in two consecutive **V1-V3** leads.

A.2. Primer Quiz

ECG Interpretation Study
Quiz

Subject ID: _____
Date: _____

Please match each heart pattern to its corresponding description

___ 1. Anterior STEMI

___ 2. Left Axis Deviation

___ 3. Right Bundle Branch Block

___ 4. Inferior STEMI

___ 5. Right Axis Deviation

___ 6. Normal

___ 7. Old Inferior MI

A. No abnormalities

B. QRS > 0.12s and rsR' or rSR' (i.e. "rabbit ears") in V1 & V2; deep reciprocal S waves in left lateral leads

C. QRS in I is negative and aVF is positive, or QRS is evenly divided in I with III more positive than aVF.

D. ST elevation >2 mm in two consecutive V1-V3 leads

E. Significant Q's (0.04 sec & > ¼ the height of R) in at least two of II, III, and aVF; no ST elevation; may have nonspecific T waves. (May result in Left Axis Deviation pattern)

F. R net positive in I and net negative in II and aVF

G. ST elevation >1 mm in II, III and aVF with ST depression > 1 mm in leads I and aVL.

APPENDIX B

B.1. Survey Questions for ECG Experiments 1, 3, 5

This survey was given to participants following the immediate posttest.

General Questions about the Training

These questions help us understand how effective the training was for you so we can find ways to improve it. All responses will be kept anonymous and confidential. Please respond honestly.

1. How enjoyable was the training as a whole? (On a scale from 1-6, 1 = Not at all enjoyable, 6 = very enjoyable)
2. How much did you learn from today's session? (1 = Nothing, 6 = A lot)
3. How much of what you learn today do you think you will remember a week from now? (1 = Nothing, 6 = All of it)
4. Which of these heart patterns will you be able to identify a week from now?
    a. Normal
    b. Anterior STEMI
    c. Inferior STEMI
    d. Right Bundle Branch Block
    e. Left Axis Deviation
    f. Right Axis Deviation
    g. Old Inferior MI
    h. None of these
5. Which of these heart patterns were the most difficult to identify?
    a. Normal
    b. Anterior STEMI
    c. Inferior STEMI
    d. Right Bundle Branch Block
    e. Left Axis Deviation
    f. Right Axis Deviation
    g. Old Inferior MI
    h. None of these
6. How helpful were the primer powerpoint slides? (1 = Not at all helpful, 6 = Extremely helpful)
7. How can we improve the primer?
8. How helpful was the training module (1 = Not at all helpful, 6 = Extremely helpful)
9. How motivated and engaged were you during the training? (1 = Not at all, 6 = Very much so)
10. How can we improve the training module?
    a. How can we improve the design of the module and the user experience? Ex. Buttons, *F*eedback, sounds, layout, graphics, loading speed, etc.
11. What prior experience do you have with interpreting ECGs? (if any)
12. What are your interests and background in the medical field? *Are you thinking of pursuing a medical degree? What general or specific interests do you have in heart functioning or reading ECGs?*
13. How many hours of sleep did you have last night?
    a. Less than 4 hours

b. 4-6 hours
c. 6-8 hours
d. More than 8 hours

14. Any comments about your experience in today's session? *What can we improve on? Did you experience any technical difficulties? Did we make any error in the training or in the assessments?*

Theory of Intelligence Survey (from Chiu, Hong, & Dweck, 1997)

Please indicate how much you agree with the following statements on a scale from 1-6 (1 = strongly disagree, 6 = strongly agree):

1. You have a certain amount of intelligence and you really cannot do much to change it.
2. Your intelligence is something about you that you cannot change very much.
3. You can learn new things, but you cannot really change your basic intelligence.
4. To be honest, you can't really change how intelligent you are.

Demographic Questionnaire

These questions are for demographic purposes. Your responses will be kept anonymous and confidential. Please respond honestly.

1. What is your age?
2. What is your gender?
3. What is your year in college?
    a. Freshman
    b. Sophomore
    c. Junior
    d. Senior
    e. Graduate Student
    f. Other: _____
4. What is your major?
5. How fluent are you in English? (1 = Not at all fluent, 5 = Near native or Native)
6. Which of the following apply to you? *Please select all that apply.*
    a. I took a gap year (or multiple gap years) before college
    b. I am a transfer student
    c. I consider myself an older adult/mature student
    d. I am an international student
    e. I am an immigrant or refugee
    f. My parent(s) is/are immigrant(s) or refugee(s)
    g. I am a first-generation college student
    h. None of these apply to me

(Experiment 4 ONLY) In this study, we compare the effectiveness among three versions of the training module. They only differ in how the questions appear on the screen. Which version do you think would produce the highest posttest score?

a. Version 1: Each question shows an unknown ECG and asks you to classify it.

b. Version 2: Each question shows 2 ECGs (one Normal and one unknown), and asks you about the unknown ECG.

c. Version 3: An equal mixture of Version 1 and Version 2.
d. They are equally effective.

B.2 Survey Questions for AlgGeo Experiments 2, 4, 6

There were two surveys. Survey 1 was given after the immediate posttest, and Survey 2 was given after the delayed test.

SURVEY 1
These questions help us understand your experience in the training so we can find ways to improve it. All responses will be kept anonymous. Please respond honestly. Your answers will not affect payment.

1. How enjoyable was the training? (1 = Not at all, 6 = Very much)
2. How helpful was the training? (1 = Not at all helpful, 6 = Extremely helpful)
3. How motivated and engaged were you during the training? (1 = Not at all, 6 = Very much so)
4. What else were you doing during the study? *This is important because we want to know whether your speed on each question was an accurate measurement of how fast you were. If you were not doing anything else, please write NOTHING.*
5. How well did you know about Sine and Exponential transformations BEFORE this training? (1 = Not at all, 6 = Very well)
6. How well do you feel you know about Sine and Exponential transformations NOW? (1 = Not at all, 6 = Very well)
7. Which were the most difficult type of transformation to recognize? *Select up to 5.*
    a. sin(x +- 2)
    b. sin(x) +- 2
    c. sin(x */ 2)
    d. sin(x) */ 2
    e. exp(x +- 2)
    f. exp(x) +- 2
    g. exp(x */2)
    h. exp(x) */ 2
    i. All of them were equally difficult
8. How much of what you learned today will you remember a week from now? (1 = Not at all, 6 = All of it)
9. What were your strategies during the training? Are they different from your strategies in the posttest? If so, how?
10. How can we improve the training module?
11. Did you experience any technical difficulty? If so, please describe.
12. Did you consult any outside resources to help you with the materials at any point during the study? (ex: use a calculator, ask someone for help, looked up the answers, etc.). *If YES, please indicate the resource(s) you used. If NO, please write NO.*
13. Did you take breaks during the study? If so, please share what you did during the break(s). *We look for things that may affect your learning, such as, did you take a nap? a walk? had a snack? etc.*

14. Any other comments about the training or your experience in this study so far?

Describe the transformations (Experiment 6 ONLY)

1. From $y = sin(x)$, how do we get $y = sin(x + 2)$?
    a. Shift to the left 2 units
    b. Shift to the right 2 units
    c. Shift upward 2 units
    d. Shift downward 2 units
    e. Expand horizontally 2 times
    f. Compress horizontally 2 times
    g. Expand vertically 2 times
    h. Compress vertically 2 times
2. From $y = sin(x)$, how do we get $y = sin(x) / 2$? (Same answer choice as above)
3. From $y = exp(x)$, how do we get $y = exp(2x)$? (Same answer choice as above)
4. From $y = exp(x)$, how do we get $y = exp(x) - 2$? (Same answer choice as above)

Theory of Intelligence Survey (same as survey used in ECG)

Student Beliefs about Mathematics Survey (Modified from Kaya, 2008)

How true are these statements to you?

1. "I like math." (1 = Not at all true, 4 = Very true)
2. "I learn math by understanding the underlying logical principles, not by memorizing rules." (1 = Not at all true, 4 = Very true)
3. "I feel nervous when I do math because I think it's too hard" (1 = Almost never, 4 = Almost all of the time)

Demographic Questionnaire

These questions are for demographic purposes. Your responses will be kept anonymous and confidential. Please respond honestly.

1. What is your age?
2. What is your gender?
3. Which US state do you live in?
4. What is your job?
5. How much math knowledge does your current job require? (1 = None, 6 = It's all math!)
6. When was the last time you took a math class or a class that involved a lot of math?
    a. Currently
    b. Within a year
    c. 1 to 2 years ago
    d. More than 2 years ago
7. What was the last math class (or a class that involved a lot of math) that you took?
8. What is the highest degree or level of school you have completed?
    a. Up to 8th grade
    b. Some high school, no diploma
    c. High school
    d. Some college
    e. Trade/technical/vocational training

    f. Bachelor's degree
    g. Graduate or professional degree
    h. Prefer not to answer
9. How fluent are you in English? (1 = Not at all fluent, 5 = Near native or Native)
10. What is your ethnicity or race? *Select all that apply*
    a. White
    b. Hispanic or Latino/a
    c. Black or African American
    d. Native American/Alaska Native
    e. Asian/Pacific Islander
    f. Prefer not to say
    g. Other: _____
11. How many hours of sleep did you have last night?
    a. Less than 4 hours
    b. 4-6 hours
    c. 6-8 hours
    d. More than 8 hours
12. Why did you decide to do this study? *Select all that apply*
    a. The pay is good
    b. It seems fun or interesting
    c. I like to contribute to research
    d. I wanted to learn some math

SURVEY 2

These questions help us understand your experience to better evaluate the effectiveness of the training. All responses will be kept anonymous and confidential. Please read each question carefully and respond honestly. Your responses will not affect payment.

1. How did you find the delayed test questions today? (1 = Extremely easy, 6 = Extremely hard)
2. How motivated were you to get each question correctly and quickly in this assessment? (1 = Not at all, 6 = Very much so)
3. How well do you think you know about Sine and Exponential transformations now? (1 = Not at all, 6 = Very well)
4. How well do you think you know about Cosine and Log transformations now? (1 = Not at all, 6 = Very well)
5. What was your strategy during the delayed test?
6. What else were you doing during the study? *This is important because we want to know whether your speed on each question was an accurate measurement of how fast you were. If you were not doing anything else, please write NOTHING.*
    a. If you answered "YES" in the previous question, please explain what you did.
7. Did you experience any technical difficulty? If so, please describe.
8. Did you review, study, or look up the materials covered in this study during this past week? If YES, please indicate how. If NO, please write NO.
9. Did you consult any outside resources to help you with the materials at any point during this Part 2 of the study? If YES, please indicate the resource(s) you used. If NO, please write NO.

10. How many hours of sleep did you have last night?
    a. Less than 4 hours
    b. 4-6 hours
    c. 6-8 hours
    d. More than 8 hours

Describe the transformations (Experiment 6 ONLY) – Same as in Survey 1

Any other comments?

APPENDIX C

C.1. Experiment 1 – Results from All Participants

Here we report data from 81 undergraduates (61 Female, age mean = 19.78) who participated in the Experiment 1, of whom 67 completed the modules. All assumptions were met.

Trial Efficiency

After controlling for the effect of pretest accuracy, the 2 phase (pre-post, pre-delayed) x 3 condition (*active, passive, passive-active*) ANCOVA on trial efficiency confirmed main effects of condition on efficiency, $F(2, 77) = 6.10$, $p < .01$, $\eta^2_p = .14$. There were no differences between the *active* and *passive* conditions on either efficiency scores (p's > .05), nor between the *active* and the *passive-active* conditions (p's > .05), but the *passive-active* condition was significantly more efficient than the *passive* condition on both the pre-post efficiency score, $M = .003$ vs. .002, respectively, $t(52) = 2.68$, $p < .05$, $d = .73$, and the pre-delayed post efficiency score, .002 vs. .001, respectively, $t(52) = 2.58$, $p < .05$, $d = .70$.

The main effect of phase was marginally significant, $F(1, 77) = 2.83$, $p = .10$, $\eta^2_p = .04$, reflecting the drop in accuracy from immediate posttest to delayed test seen in most participants. The pre-post efficiency ($M = .002$, $SD = .002$) was marginally higher than the delayed test efficiency ($M = .001$, $SD = .002$), paired-$t(80) = 6.88$, $p < .001$, $d = .69$.

The covariate, pretest accuracy, was significantly related to the efficiency scores, $F(1,77) = 24.66$, $p < .001$, $\eta^2_p = .24$. Predictably, pretest accuracy is strongly and negatively correlated with the pre-post efficiency, $r(81) = -.35$, $p < .001$, as well as pre-delayed post efficiency, $r(81) = -.48$, $p < .001$.

Figure C.1a shows the average accuracy at each test phase, and *Figure C.1b* shows the trial efficiency by condition. As expected, participants showed substantial learning gains from

pretest to immediate posttest and retained much of their learning at delayed test, regardless of condition. Participants were able to interpret ECGs they had never seen before and to do so with improved speed. The *passive-active* condition produced the greatest learning gain with the fewest training trials. The *active* condition also produced greater learning gains than the *passive* condition. *Table C.1* contains the descriptive statistics from the training for each condition.

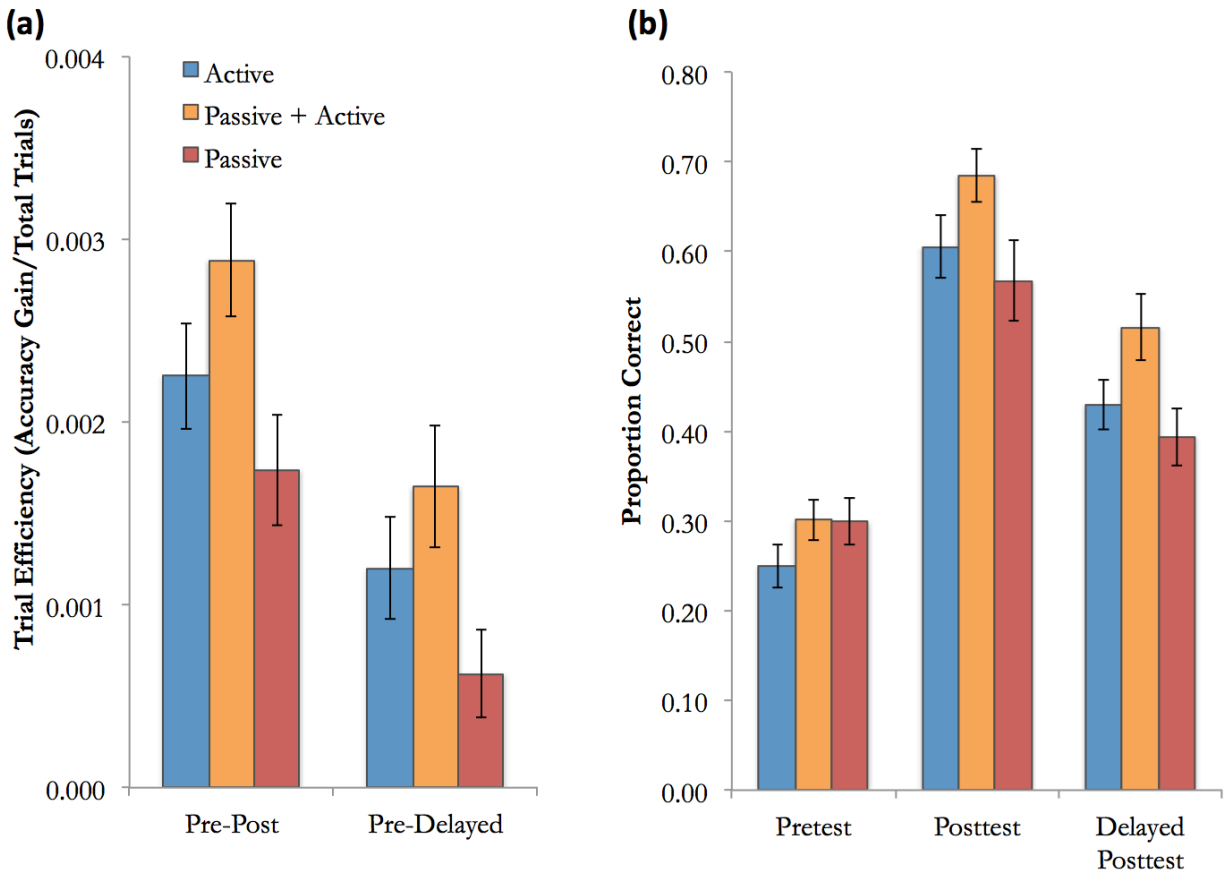


Figure C.1. (a) Trial efficiency and (b) Accuracy by conditions. Error bars ± 1SE.

Accuracy

Accuracy Gain. We analyzed accuracy gains (posttests minus pretest) in a 2 phase (pre-post, pre-delayed) x 3 condition (*active*, *passive*, *passive-active*) repeated-measures ANCOVA with pretest accuracy as the covariate. The covariate, pretest accuracy, was significantly related

to the posttest gains, $F(1,77) = 35.42$, $p < .001$, $\eta^2_p = .32$. Indeed, better pretests predicted less improvement at immediate posttest, $r = -.38$, $p < .001$, and delayed test, $r = -.52$, $p < .001$, suggesting that pretest variations were largely due to chance. After controlling for the effect of the pretest, there was a reliable effect of condition, $F(2, 77) = 4.24$, $p < .05$, $\eta^2_p = .10$. The *active* and *passive-active* conditions produced higher gains than the *passive* condition (27% and 30% vs. 18%; $t(52) = 1.75$, $p = .09$, $d = .48$, $t(52) = 2.31$, $p < .05$, $d = .63$, respectively. There were no reliable differences in accuracy gains between the *passive-active* and *active* conditions and no significant interactions (p's > .10).

There was a statistically significant main effect of phase, $F(1, 77) = 5.49$, $p < .05$, $\eta^2_p = .07$. The pre-post accuracy gain was reliably higher than the pre-delayed gain (34% vs. 16%, respectively, $d = .83$).

Raw Accuracy. We also compared raw accuracy across groups. A 3 phase (pre, post, delayed test) x 3 condition ANOVA confirmed a main effect of condition, $F(2,78) = 3.90$, $p = .02$, $\eta^2_p = .09$. The *passive-activ*e condition outperformed both the *active*, $t(52) = 2.56$, $p = .01$, $d = .70$, and *passive* conditions, $t(52) = 2.41$, $p = .02$, $d = .67$, on overall accuracy. *Active* and *passive* did not differ reliably, $p > .10$. There was no significant phase x condition interaction, $p > .10$.

Response Times

Generally, participants became faster at arriving at the correct answers at immediate posttest and delayed test. However, there were no reliable effects of condition or phase (pre-post vs. pre-delayed post), p's > .05.

Progression of Learning

Table C.1 shows the average training performance by condition. *Figure C.2* shows the average accuracy over the first 17 training blocks for the *active* and *passive-active* conditions. The *passive-active* group performed consistently better than the *active* group, $t(52) = 2.96$, $p < .001$, $d = .80$, after 3 blocks, $t(52) = 2.95$, $p < .01$, $d = .80$. This result suggests that initial passive exposure speeds learning relative to starting with active classification, despite the similar number of learning trials in the passive portion and the first active trial block. In the first few blocks, the abrupt change from passive to active introduced similar error rates as those in the *active* condition. However, after the first few blocks, as we expected, those in the *passive-active* group made fewer errors, presumably because the initial passive learning freed them from the performance demands, and they could concentrate on deepening their understanding of the categories. These gains appeared to be preserved through the course of learning and in posttests.

Condition	Total Trials Completed	Minutes on Training	Training Accuracy	Percent Mastery	Proportion reached 100% mastery
Active	167.52 (11.82)	43.96 (3.37)	.49 (.02)	87.3 (6.12)	23/27
Passive-active	137.78 (6.69)	37.70 (2.68)	.57 (.02)	89.9 (4.94)	23/27
Passive	159.74 (8.64)	47.91 (2.40)	--	--	--

Table C.1. Average training performance across the three experimental groups (Standard errors in parentheses). Both passive and active trials were included in total trials completed for the *passive-active* condition.

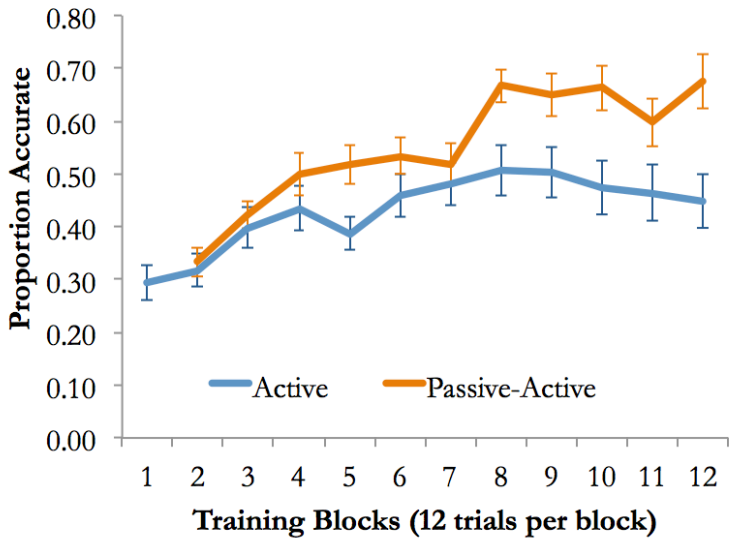


Figure C.2. Mean accuracy across training blocks. The *passive-active* group received 14 *passive* trials at block 1.

Self-Report Ratings

On the survey, participants differed marginally in how they responded to "How enjoyable was the training as a whole, on a scale from 1-6 (1 = not at all enjoyable, 6 = very enjoyable)"[15], $F(2,70) = 2.99$, $p = .057$, $\eta^2 = .08$. The *passive-active* PALM was found to be reliably more enjoyable ($M = 4.42$, $SD = 1.25$) than both the *passive* PALM ($M = 3.72$, $SD = 1.17$), $t(47) = 2.01$, $p = .05$, $d = .58$, and the *active* PALM ($M = 3.58$, $SD = 1.38$), $t(46) = 2.19$, $p < .05$, $d = .64$.

Interestingly, participants in the *passive* training condition gave themselves marginally higher ratings to "On a scale from 1-6 (1 = nothing, 6 = all of it), how much of what you learn today do you think you will remember a week from now" (*Passive,* $M = 3.68$, $SD = .69$, vs. *passive-active,* $M = 3.25$, $SD = 1.03$, $t(47) = 1.72$, $p = .092$, $d = .49$, and vs. *active,* $M = 3.25$, *SD*

[15] The survey was implemented shortly after data collection began, so we did not have responses from the first 8 participants.

= 1.07, $t(47) = 1.68$, $p = .10$, $d = .48$). This observation is consistent with research on the illusion of competency in judgments of learning (Castel, McCabe, Roediger, 2007). Without having to select answers and receive feedback during the training, the *passive* group experienced an easier time during the training, which likely led to the misbelief that they would have better memory of the materials on the delayed test. Active classification practice then, may also support effective self-regulated learning.

When asked, "Which heart pattern(s) were the most difficult to identify?" (participants were allowed to pick up to 4 or to choose "All 7 patterns were equally difficult"), most participants rated Normal to be most difficult (41/81 participants), Followed by LAD (38/81), OldInfMI (36/81), RAD (32/81) and InfSTEMI (30/81), and least difficult were AntSTEMI (16/81) and RBBB (10/81). They gave the opposite pattern of responses to, "Which heart pattern(s) will you remember a week from now?" Most people indicated that they would remember RBBB (67/73) and Anterior STEMI (51/73), followed by Inferior STEMI (40/73), RAD (39/73), OldInfMI (38/73), LAD (31/73), and Normal (20/73).

We also asked participants to respond to a theory of intelligence short survey, and to a question about their educational hardship. Having a growth mindset and having had prior educational hardship may contribute to participants' likelihood to persevere in the module. We averaged participants' six-point Likert scale across the four theory of intelligence statements. We categorized participants as either "fixed" or "growth" theorists, as has been the practice in the prior literature (e.g., Blackwell, Trzesniewski, & Dweck, 2007; Miele & Molden, 2010), and refer to this variable as categorical theory of intelligence: Those scoring an average above 3.5 were classified as fixed theorists, while those scoring an average below 3.5 were classified as growth theorists. 65.5% of the participants had a growth mindset. There were no differences

between the mindset groups on any of the dependent variables (p's > .05). There were also no correlations between having had experienced common educational hardships and any of the dependent variables (p's > .05).

C.2. Experiment 1 - Extra Analyses and Detailed Results

These analyses were from data of participants who have reached learning criterion.

Efficiency by Time

After controlling for the effect of pretest accuracy, there was a reliable main effect of condition, $F(2, 65) = 10.04$, $p < .001$, $\eta^2_p = .24$. There were no reliable differences between *active* and *passive-active* groups in the average efficiency ($p = .11$), but both of the *active* and *passive-active* groups had better efficiency than the *passive* group with medium to large effect sizes (.007 and .010 vs. .004, respectively, $t(44) = 2.34$, $p < .05$, $d = .69$, and $t(44) = 4.41$, $p < .001$, $d = 1.01$. The drop in efficiency from immediate posttest to delayed test was also reliable, $F(1, 65) = 4.12$, $p < .05$, $\eta^2_p = .06$. There was no phase x condition interaction, $F(2,65) = .87$, $p = .87$, $\eta^2_p = .004$. Pretest accuracy was also significantly related to time efficiency scores, $F(1, 65) = 26.58$, $p < .001$, $\eta^2_p = .29$, with pretest accuracy negatively correlated with both pre-post efficiency, $r(69) = -.37$, $p < .001$ and pre-delayed post efficiency, $r(69) = -.52$, $p < .001$.

We also analyzed time efficiency uncorrected for pretest variations. Across both posttests, the *passive-active* condition outperformed the *active* condition with medium and large effect sizes (.019 vs. .014 and .011, $t(44) = 2.15$, $p < .05$, $d = .62$, and $t(44) = 4.41$, $p < .001$, $d = 1.30$, respectively). Although there was no difference between the *active* and *passive* condition

on trial efficiency, the *active* condition produced higher time efficiency uncorrected for pretest accuracy with a large effect size, $t(44) = 2.73$, $p < .01$, $d = .82$.

Fluent Accuracy

Across conditions, the improvements in fluent accuracy had very large effect sizes (pretest to immediate posttest: 19% to 57%, $t(68) = 15.98$, $p < .001$, $d = 2.63$, and pretest to delayed test, 19% to 42%, $t(68) = 9.89$, $p < .001$, $d = 1.65$). The drop between immediate posttest and delayed test was also reliable, $t(68) = 7.47$, $p < .001$, $d = .92$.

Response Times on Correct Answers (RTc)

Generally, participants became faster at arriving at the correct answers at immediate posttest and delayed test. As participants improved in accuracy, they also improved in speed. At pretest, participants took about 12.07 seconds per correct response, and at immediate posttest 8.91 seconds and delayed test 8.90 seconds (pre vs. post, $t(68) = 6.60$, $p < .001$, $d = 1.00$, and pre vs. delayed, $t(68) = 6.59$, $p < .001$, $d = .98$). There were no reliable effects of condition or of phase on RTc gain, p's > .05.

APPENDIX D

D.1. Experiment 2 - Demographic Data

We lost survey data from one participant, so the following survey data were based on 74 participants rather than 75.

Location

Participants came from 34 US states.

Math Background

Most participants have not had recently taken a mathematics class or a class with a heavy emphasis on mathematics such as physics. The majority (88%) has had the last math class over 2 years ago, 5.3% 1-2 years ago, 4% within a year ago, and only 1.3% are currently in a math class.

In response to "How much math is involved in your current job on a scale from 1-6 (1 = not at all, 6 = it's all math)", 96% of participants chose 1-4, indicating that their jobs did not require much math.

Education Level

40% Bachelor's degree, 34.7% some college, 13.3% has a graduate or professional degree, 5.3% has a high school degree, 4% has trade, technical, or vocational training degree, and 1.33% has some high school (no high school diploma).

Ethnicity

78% White, 6% Asian/Pacific Islander, 4% Hispanic or Latino/a, 4% Black or African American, 6% Mixed, 2% did not report.

English Fluency

On a scale from 1-5, 1 = not at all fluent, 5 = native or near native, 95.9% of participants rated 5, the other 4.1% rated 4.

Reason for Participation

Participants could select more than one reason: 55/75 participants chose "The pay is good", 53/75 chose "It seems fun or interesting", 31/75 chose "I wanted to learn some math" and a math refresher, 35/75 chose "I like to contribute to research".

Attitude about Mathematics

Not central to the study was a short scale of participants' general attitude about mathematics. These questions appear in the *Appendix B.2*. We reverse-coded participants' response to this statement "I learn math by understanding the underlying logical principles, not by memorizing rules," and averaged their ratings on the 3 questions. The average rating was 2.79 (*SD* = .70). Higher ratings suggest a more positive attitude toward math. As we expected, the higher the rating, the better participants performed on the immediate posttest, $r(74) = .43$, $p < .001$. This was true for all of the assessment types: $r(74) = .49$, $p < .001$ for trained items (TI), $r(74) = .28$, $p < .05$ for trained functions, novel items (TF/NI), $r(74) = .24$, $p < .05$ for untrained functions (UF) and $r(74) = .22$, $p = .06$ for combination functions (CF). This in turn meant higher pre-post accuracy gain, $r(74) = .37$, $p < .01$, and higher pre-post efficiency scores calculated by trial, $r(74) = .41$, $p < .001$, and by time, $r(74) = .36$, $p < .01$.

Positive math attitudes only marginally correlated with delayed test accuracy, $r(74) = .23$, $p = .05$, and significantly correlated with the trained items (TI) accuracy, $r(74) = .20$, $p = .09$.

Higher attitude score also was correlated with the accuracy and fluent accuracy during the training for those in the active and passive-active conditions, $r(49) = .29$, $p < .05$ for accuracy, and $r(49) = .35$, $p < .05$ for fluent accuracy. It was also marginally negatively correlated to the total number of training trials completed, $r(49) = -.26$, $p = .07$.

D.2. Experiment 2 – Details of Reported Results

Efficiency by Trial

Trained Items (TI)

The 2 phase x 3 condition ANCOVA confirmed a significant main effect of condition, $F(2,71) = 5.09$, $p = .009$, $\eta^2_p = .13$. Both the *passive-active* ($M = .0022$, $SD = .0023$) and the *active* ($M = .0017$, $SD = .0021$) conditions produced higher efficiency than the *passive* ($M = .0007$, $SD = .0013$) condition with medium effect sizes, $t(48) = 2.71$, $p = .009$, $d = .80$, and $t(48) = 1.98$, $p = .05$, $d = .57$, respectively. There were no differences between the *passive-active* and the *active* conditions, $p > .10$, and no phase x condition interaction, p's $> .10$.

There was no main effect of phase, $F(1,71) = .03$, $p > .10$. This reflected good retention on trained items after a week delay. The pretest TI correlated strongly and negatively with the pre-post trial efficiency, $r(75) = -.35$, $p < .01$, and with the pre-delayed trial efficiency on the same items, $r(75) = -.49$, $p < .01$, suggesting that variations at pretest were likely due to chance.

Trained Functions/Novel Items (TF/NI)

After controlling for differences in pretest TF/NI accuracy, a 2 x 3 ANCOVA with pretest TF/NI as the covariate confirmed a main effect of condition, $F(2,71) = 4.39$, $p = .02$, $\eta^2_p = .11$. Indeed, the *passive-active* group ($M = .0021$, $SD = .0020$) surpassed the *passive* group ($M = .0006$, $SD = .0011$) on TF/NI trial efficiency with a large effect size, $t(48) = 3.28$, $p = .002$, $d = .95$. The *passive-active* group also had marginally higher efficiency than the *active* group ($M = .0011$, $SD = .0020$) with a medium effect size, $t(48) = 1.81$, $p = .08$, $d = .53$. There were no main effect of phase and phase x condition interaction, p's > .10.

There was a main effect of pretest TF/NI accuracy, $F(1,71) = 22.56$, $p < .001$, $\eta^2_p = .24$. The pretest TF/NI accuracy strongly and negatively correlated with both the pre-post TF/NI trial efficiency, $r(75) = -.38$, $p < .001$, and the pre-delayed TF/NI trial efficiency, $r(75) = -.52$, $p < .001$.

Untrained Functions (UF)

There were no condition nor phase differences in trial efficiencies, p's > .10.

Combination Functions (CF)

There were no condition nor phase differences in trial efficiencies, p's > .10.

Efficiency by Time

Overall

The 2x3 ANOVA confirmed a main effect of phase, $F(1,72) = 18.41$, $p < .001$, $\eta^2_p = .20$. This reflected how most participants performed higher on the immediate posttest than after a one-week delay. The pre-post trial efficiency ($M = .006$, $SD = .006$) was reliably higher than the

pre-delayed efficiency but with a small effect size (M = .0036, SD = .005), $t(74) = 4.32$, $p < .001$, d = .45.

There was a main effect of condition, $F(2,72) = 5.74$, $p < .01$, $\eta^2_p = .14$. Both the *passive-active* condition (M = .007, SD = .005) and the *active* condition (M = .005, SD = .005) produced higher time efficiency than the passive condition (M = .003, SD = .002) with medium and large effect sizes, $t(48) = 3.57$, $p = .001$, d = .99, and $t(48) = 2.06$, $p = .04$, d = .58, respectively. There was no phase x condition interaction, $F(2,72) = .47$, $p < .10$, $\eta^2_p = .01$.

Trained Items

Time efficiency on TI showed the same pattern as when calculated with trial efficiency. After controlling for the differences in pretest TF/NI accuracy, there was no main effect of phase, $p > .10$, but there was a main effect of condition, $F(2,71) = 5.64$, $p = .005$, $\eta^2_p = .14$, and no phase x condition interaction, $p > .10$. Both of the *passive-active* (M = .008, SD = .009) and the *active* (M = .008, SD = .01) groups produced higher efficiency than the *passive* group (M = .003, SD = .004), $t(48) = 2.88$, $p = .006$, d = .81, and $t(48) = 2.68$, $p = .01$, d = .76, respectively. There was no difference between the *passive-active* and the *active* condition, $p > .10$.

Trained Functions, Novel Items

After controlling for the differences in pretest TF/NI accuracy, there was a main effect of condition, $F(2,71) = 6.13$, $p = .004$, $\eta^2_p = .15$. Indeed, the *passive-active* group (M = .008, SD = .007) surpassed the *passive* group (M = .002, SD = .003) on TF/NI trial efficiency with a large effect size, $t(48) = 3.86$, $p < .001$, d = 1.10. The *passive-active* group also had marginally higher efficiency than the *active* group (M = .005, SD = .007), with medium effect size, $t(48) = 1.74$, p = .09, d = .50. Interestingly, the *active* group also had marginally higher efficiency than the *passive*

group, thought with a small effect size, $t(48) = 1.68$, $p < .10$, $d = .48$. This was not found with trial efficiency. There were no main effect of phase and no phase x condition interaction, p's > .10.

There was a main effect of pretest TF/NI accuracy, $F(1,71) = 25.27$, $p < .001$, $\eta^2_p = .26$. The pretest TF/NI accuracy strongly and negatively correlated with both the pre-post TF/NI trial efficiency, $r(75) = -.35$, $p < .01$, and the pre-delayed TF/NI trial efficiency, $r(75) = -.51$, $p < .001$.

Untrained Functions

After controlling for the differences in pretest UF accuracy, although there were marginally significant main effect of phase, $F(1,71) = 3.24$, $p = .08$, $\eta^2_p = .04$ and a significant main effect of condition, $F(2,71) = 3.83$, $p = .03$, $\eta^2_p = .10$, there were no reliable differences between any of the three conditions, p's > .10. There was no phase x condition interaction, $p > .10$.

There was a main effect of pretest TF/NI accuracy, $F(1,71) = 40.80$, $p < .001$, $\eta^2_p = .37$. The pretest TF/NI accuracy strongly and negatively correlated with both the pre-post TF/NI trial efficiency, $r(75) = -.51$, $p < .001$, and the pre-delayed TF/NI trial efficiency, $r(75) = -.41$, $p < .001$.

Combination Functions

There were no condition nor phase differences, p's > .10.

Accuracy

Trained Items (TI)

Exponential TI

Raw Accuracy. All groups showed strong learning gains and retention on Exponential TIs, $F(2,144) = 32.14$, $p < .001$, $\eta^2_p = .31$. There was a marginally significant main effect of condition, $F(2,72) = 3.05$, $p = .05$, $\eta^2_p = .01$. The *passive-active* group ($M = .47$, $SD = .16$) did better than both the *active*, $t(48) = 2.15$, $p = .04$, $d = .59$, and the *passive* groups ($t(48) = 2.18$, $p = .03$, $d = .65$, on Exponential TI, with medium effect sizes. *Active* ($M = .37$, $SD = .18$) did not differ from *passive* ($M = .37$, $SD = .15$), $p > .10$, and there was no phase x condition interaction, $F(2,144) = 1.45$, $p = .22$, $\eta^2_p = .04$.

Accuracy gain. There was a marginally significant phase x condition interaction, $F(1, 71) = 2.83$, $p = .07$, $\eta^2_p = .07$. The *passive-active* group ($M = .20$, $SD = .38$) did marginally better than the *active* group ($M = .05$, $SD = .25$) on the pre-delayed test gain, $t(48) = 1.67$, $p = .10$, $d = .47$. There were no other condition differences on Exponential TI accuracy gains, p's $> .10$.

Sine TI

Raw Accuracy. Although all groups showed strong learning gains and retention with Sine TI, but there were no group differences in learning gains on Sine TI ($p > .10$), no phase x condition interaction, $F(2,144) = 1.81$, $p = .13$, $\eta^2_p = .05$, and no main effect of condition, $F(2,72) = 1.18$, $p = .31$, $\eta^2_p = .03$.

Accuracy gain. There was a marginal main effect of condition on accuracy gains, $F(2,71) = 2.68$, $p = .08$, $\eta^2_p = .07$. Both the *active* ($M = .23$, $SD = .31$) and the *passive-active* conditions ($M = .22$, $SD = .26$) outperformed the *passive* condition ($M = .05$, $SD = .25$), $t(48) = 2.28$, $p = .03$, $d = .64$, and $t(48) = 2.45$, $p = .02$, $d = .67$. There was no reliable difference between the *passive-active* and *active* conditions, $p > .10$.

Trained Functions (TF/NIs)

Sine TF/NI

Figure D.1a shows the mean accuracy on Sine TF/NI. There were no condition differences on Sine TF/NI accuracy at any phase, p's > .10, but across conditions, participants improved significantly on Sine TF/NI items, $F(2,144) = 12.78$, $p < .001$, $\eta^2_p = .15$, from pretest ($M = .44$, $SD = .22$) to immediate posttest ($M = .62$, $SD = .24$), $t(74) = 5.28$, $p < .001$, $d = .78$, and to delayed test ($M = .55$, $SD = .24$), $t(74) = 2.81$, $p < .01$, $d = .48$. The difference between immediate posttest and delayed test was also statistically significant but with a small effect size, $t(74) = 2.20$, $p < .05$, $d = .29$.

Exponential TF/NI

Figure D.2a shows the mean accuracy on Exponential TF/NI. Similarly, regardless of condition, participants improved significantly on Exponential TF/NI items after the training, $F(2,144) = 28.83$, $p < .001$, $\eta^2_p = .29$, from pretest ($M = .31$, $SD = .20$) to immediate posttest ($M = .54$, $SD = .24$), $t(74) = 7.15$, $p < .001$, $d = 1.04$, and to delayed test ($M = .45$, $SD = .25$), $t(74) = 4.02$, $p < .001$, $d = .62$. The difference between immediate posttest and delayed test had a small effect size, $t(74) = 3.15$, $p < .01$, $d = .37$.

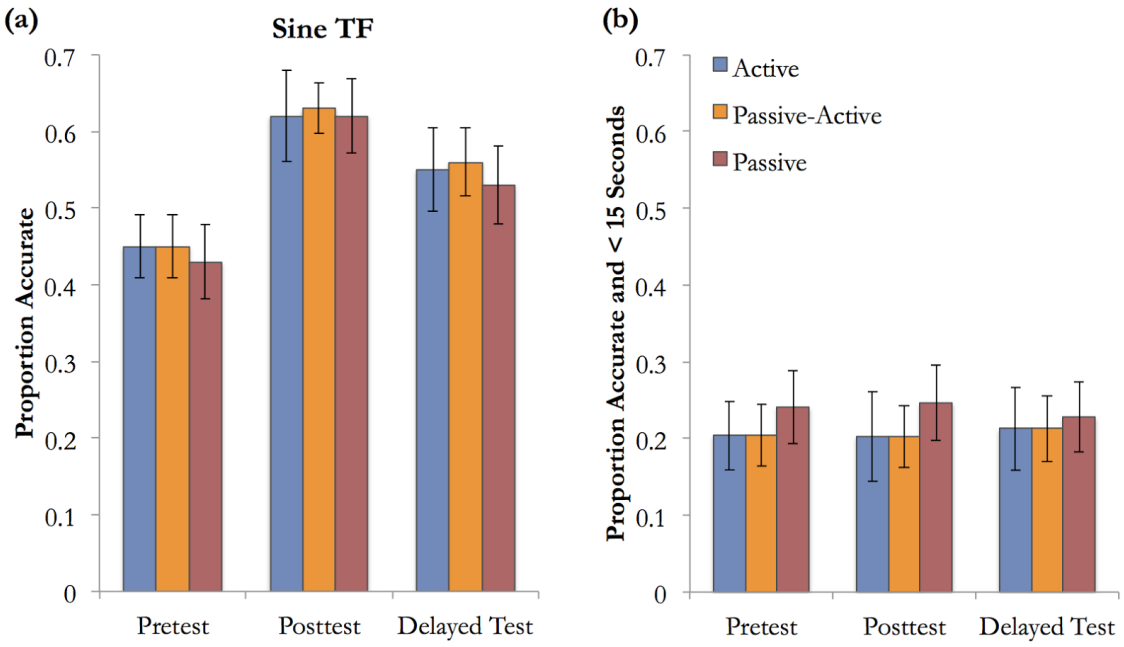


Figure D.1. Average (a) accuracy and (b) fluent accuracy on Sine TF/NI items.

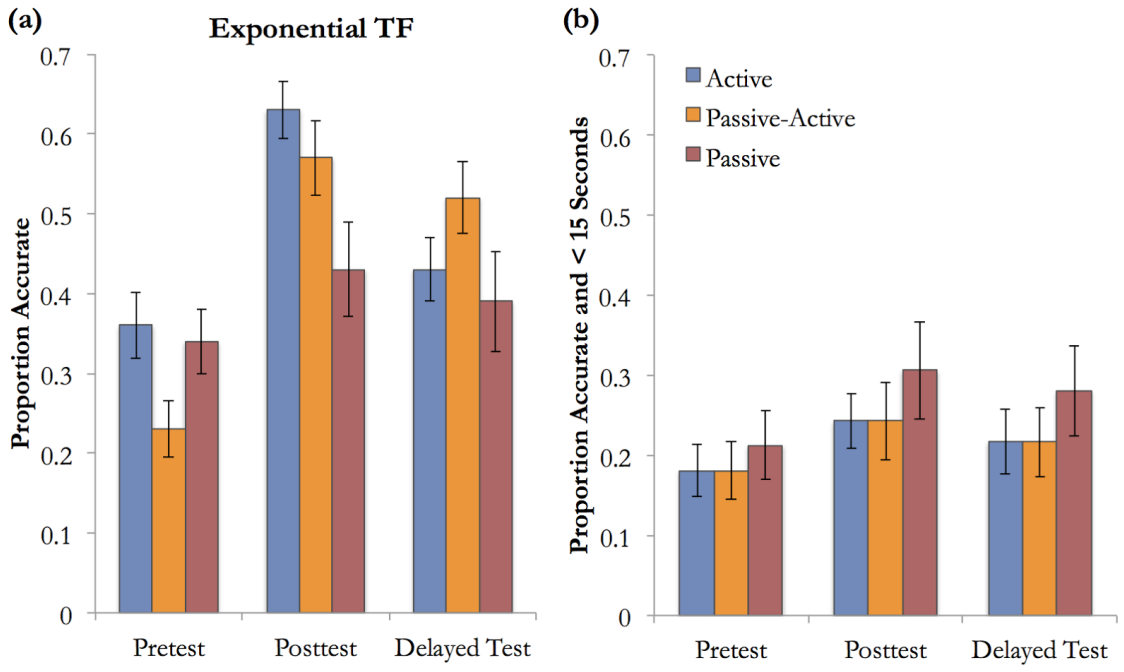


Figure D.2. Average (a) accuracy and (b) fluent accuracy on Exponential TF/NI items.

Unlike Sine TF/NI, however, there was a significant phase x condition interaction, $F(4,144) = 4.94$, $p = .001$, $\eta^2_p = .12$. At pretest, there were condition differences: the *passive* group ($M = .34$, $SD = .20$) and the *active* group ($M = .36$, $SD = .21$) started out higher than the *passive-active* groups ($M = .23$, $SD = .18$), $t(48) = 2.41$, $p = .02$, $d = .58$, and $t(48) = 2.01$, $p = .046$, $d = .66$, respectively. Interestingly, however, at immediate posttest, the *active* ($M = .63$, $SD = .18$) produced significantly higher and the *passive-active* ($M = .57$, $SD = .23$) groups produced marginally higher accuracy than the *passive* group ($M = .43$, $SD = .29$), $t(48) = 2.91$, $p = .005$, $d = .83$, and $t(48) = 1.87$, $p = .07$, $d = .53$, respectively. At delayed test, the *passive-active* group ($M = .52$, $SD = .23$) had numerically higher accuracy than the *passive* group ($M = .39$, $SD = .32$), but the difference was small and marginally significant, $t(48) = 1.67$, $p = .10$, $d = .47$. There were no other differences between the other conditions, p's > .10, and no main effect of condition, $F(2, 72) = 1.64$, $p = .20$, $\eta^2_p = .04$.

Untrained Functions (UFs)

Figure D.3a shows the accuracy on UF items. Even though participants were trained on Sine and Exponential transformations, they were able to transfer what they have learned to Cosine and Logarithmic functions, regardless of their training condition. This was confirmed by a 3 x 3 ANOVA on UF accuracy. There was a main effect of phase, $F(2,144) = 19.89$, $p < .001$, $\eta^2_p = .22$. Participants improved from pretest ($M = .29$, $SD = .15$) to immediate posttest ($M = .46$, $SD = .19$), $t(74) = 5.76$, $p < .001$, $d = .99$, and to delayed test ($M = .43$, $SD = .20$), $t(74) = 5.22$, $p < .001$, $d = .79$. Both of these mean differences were with large effect sizes. Interestingly, there was no reliable drop in accuracy on these untrained functions a week later, $p > .10$. These were no condition differences and no a phase x condition interaction, p's > .10. The 2 x 3 ANCOVA confirmed the same patterns.

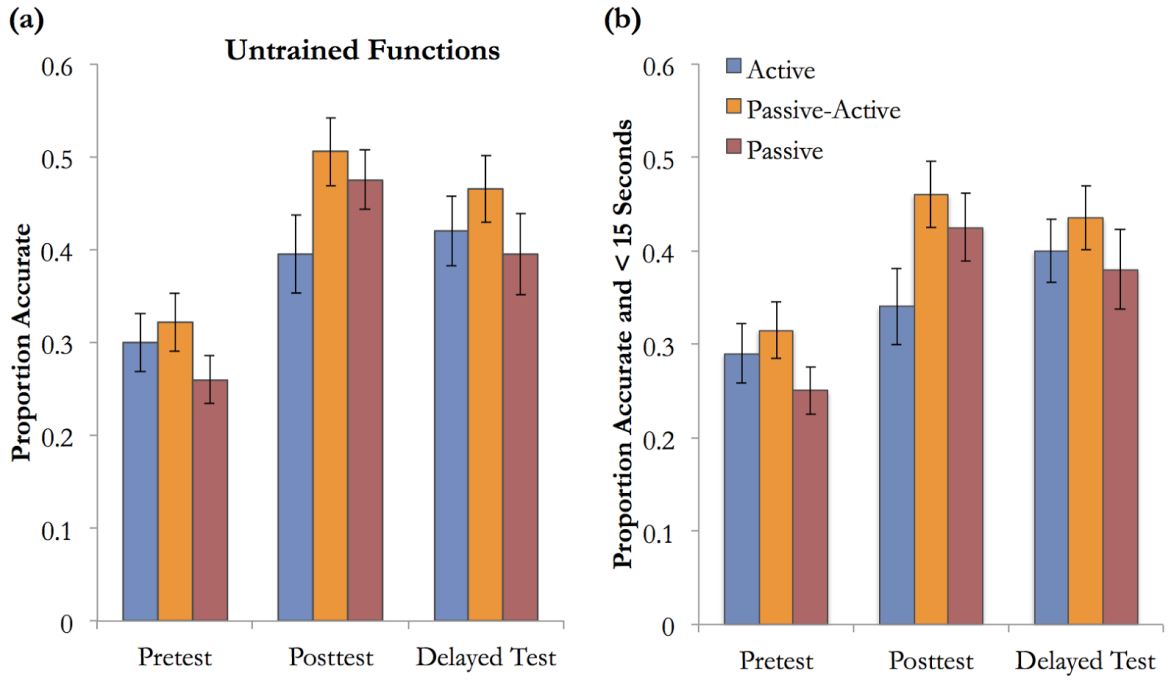


Figure D.3. Average (a) accuracy and (b) fluent accuracy on Untrained Function items.

Combination Functions (CF)

Figure D.4a shows the average accuracy on CF. Similarly, participants were able to transfer what they have learned to CF. There was a main effect of phase, $F(2,144) = 4.14$, $p < .05$, $\eta^2_p = .05$. Even on CF, participants from all conditions improved from pretest ($M = .32$, $SD = .21$) to immediate posttest ($M = .43$, $SD = .28$), $t(74) = 2.59$, $p = .01$, $d = .44$, and to delayed test ($M = .39$, $SD = .24$), $t(74) = 1.78$, $p = .08$, $d = .31$, without a reliable drop in accuracy a week later, $p > .10$. There were no main effect of condition nor phase x condition interaction, p's $> .10$.

The ANCOVA on posttest gains showed the same effects. There was a main effect of pretest accuracy, $F(1,71) = 123.58$, $p < .001$, $\eta^2_p = .64$. The pretest accuracy strongly and negatively correlated with the pre-post gain, $r(75) = -.71$, $p < .001$, and with the pre-delayed

gain, $r(75) = -.72$, $p < .001$. There were no main effect of phase, no main effect of condition, nor a phase x condition interaction, p's > .10.

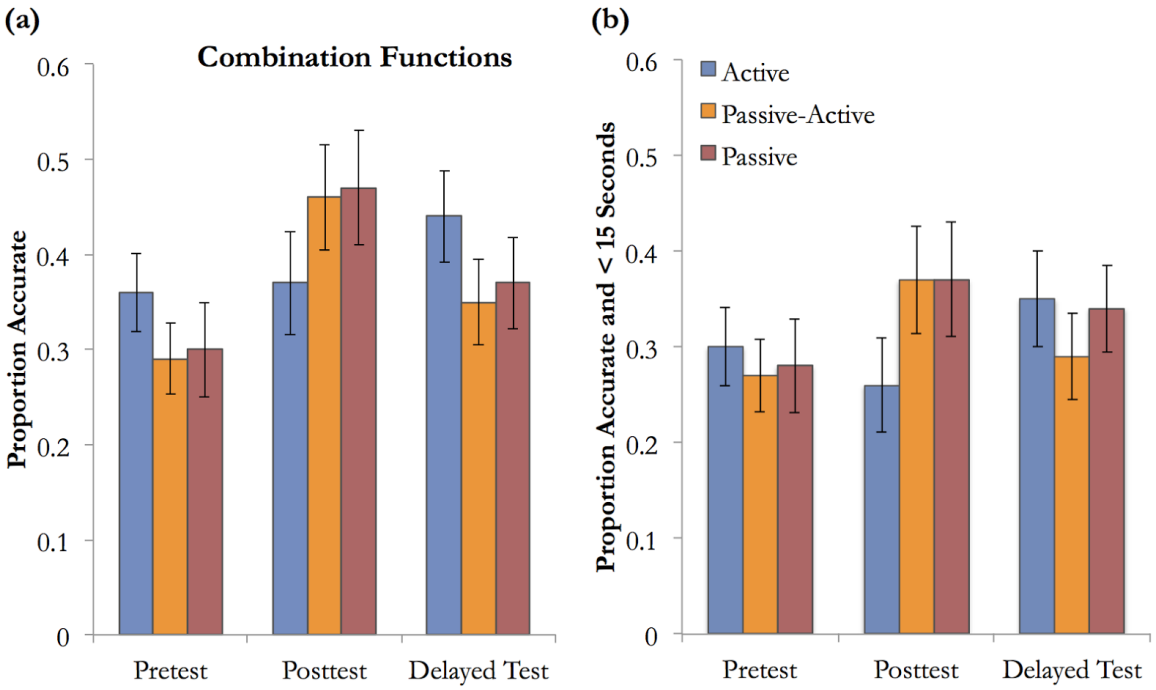


Figure D.4. Average (a) accuracy and (b) fluent accuracy on Combination Function items.

Table D.1 provides a summary of the dependent t-test statistics that examined learning gains within each condition, broken down by assessment item type. Note that all conditions improved from pretest to immediate posttest, and from pretest to delayed test overall (on all items combined), and on trained items (TI). All conditions improved from pretest to immediate posttest on trained functions, novel items (TF/NI), but not from pretest to delayed test.

		Active			*Passive-active*			*Passive*		
		t(df)	***p***	***d***	***t(df)***	***p***	***d***	***t(df)***	***p***	***d***
All Items	Pre-Post	5.30	***	1.06	7.89	***	1.58	6.80	***	1.36
	Pre-Delayed	3.90	***	0.78	5.15	***	1.03	3.24	***	0.65
TI	Pre-Post	7.27	***	1.45	6.83	***	1.37	3.76	**	0.75
	Pre-Delayed	2.26	*	0.45	3.34	**	0.67	1.71	+	0.34
Sine TI	Pre-Post	4.30	***	0.86	4.53	***	0.91	2.39	*	0.48
	Pre-Delayed	2.12	*	0.42	2.78	*	0.56	0.50	*ns*	
Exp TI	Pre-Post	6.00	***	1.20	4.62	***	0.92	3.72	**	0.74
	Pre-Delayed	1.00	*ns*		2.69	*	0.54	2.80	*	0.56
TF/NI	Pre-Post	4.99	***	1.00	6.48	***	1.30	3.47	***	0.69
	Pre-Delayed	1.78	+	0.36	4.98	***	1.00	1.62	*ns*	
Sine TF/NI	Pre-Post	2.28	*	0.46	3.84	***	0.77	3.48	***	0.70
	Pre-Delayed	1.31	*ns*		2.11	*	0.42	1.59	*ns*	
Exp TF/NI	Pre-Post	5.42	***	1.08	6.83	***	1.37	1.52	*ns*	
	Pre-Delayed	1.50	*ns*		5.07	***	1.01	0.82	*ns*	
UF	Pre-Post	1.67	*ns*		3.44	***	0.69	6.42	***	1.28
	Pre-Delayed	2.83	*	0.57	3.05	*	0.61	3.04	*	0.61
Cos	Pre-Post	1.97	+	0.39	4.06	***	0.81	5.32	***	1.06
	Pre-Delayed	1.41	*ns*		3.13	*	0.62	2.49	*	0.50
Log	Pre-Post	0.35	*ns*		0.69	*ns*		2.92	*	0.58
	Pre-Delayed	2.17	*	0.43	1.11	*ns*		1.70	+	0.34
CF	Pre-Post	0.13	*ns*		2.32	*	0.46	2.03	*	0.41
	Pre-Delayed	1.16	*ns*		1.00	*ns*		0.91	*ns*	

Table D.1. Summary of accuracy gains by assessment item type. Degrees of freedom (*df*) = 24 for all comparisons. P-value: *** denotes p < .001, ** denotes $p < .01$, * denotes $p < .05$, + denotes $p < .10$, and *ns* means not significant, $p > .10$.

Fluent Accuracy

Trained Items (TI)

Fluent Accuracy. Participants showed strong learning gains, $F(2,144) = 38.90$, $p < .001$, $\eta^2_p = .35$. Improvement from pretest ($M = .26$, $SD = .13$) to immediate posttest ($M = .49$, $SD = .20$) was robust, $t(74) = 9.29$, $p < .001$, $d = 1.36$, so was the improvement from pretest to delayed

test (M = .36, SD = .20), $t(74) = 3.63$, $p > .01$, d = .59. The drop between immediate posttest and delayed test was also reliable, $t(74) = 4.97$, $p > .001$, d = .65.

There was a marginally significant main effect of condition, $F(2,72) = 2.57$, $p = .08$, η^2_p = .07. The *active* (M = .36, SD = .12) and *passive* (M = .34, SD = .13) groups did not differ, $p >$ .10, but the *passive-active* group (M = .41, SD = .12) produced higher overall TI score than the *passive* group, $t(48) = 2.14$, $p = .04$, d = .56, and marginally higher than the active group, $t(48)$ = 1.71, $p = .09$, d = .42. There was no phase x condition interaction, $p > .10$.

Fluent Accuracy Gain. The 2 phase x 3 condition ANCOVA also showed a marginally significant main effect of condition, $F(2, 71) = 2.49$, $p = .09$, $\eta^2_p = .07$. There were no main effect of phase and no interactions, p's > .10.

Similar to accuracy, the condition differences were driven by differences in Exponential TI. There was no differences among conditions on Sine TI, but there was a marginally significant main effect of condition on Exponential TI, $F(2,72) = 3.04$, $p = .05$, $\eta^2_p = .08$. Indeed, the *passive-active* condition (M = .42, SD = .14) had higher overall fluency on Exponential TI than both the *passive* condition (M = .33, SD = .16), $t(48) = 2.25$, $p = .03$, d = .60, and the *active* condition (M = .33, SD = .18), $t(48) = 2.15$, $p = .04$, d = .56. There was no difference between the *active* and *passive* conditions, $p > .10$. There were no phase x condition interaction on Exponential TI, $p > .10$, and no reliable condition differences in fluency gains, p's > .10.

Trained Functions, Novel Items (TF/NI)

Fluent Accuracy. Across all groups, participants showed strong and persistent learning gains on TF/NI, but there were no condition differences. The 3 phase x 3 condition ANOVA supported this finding. There was a main effect of phase on TF/NI score, $F(2,144) = 31.62$, $p <$

.001, η^2_p = .31. Across all three groups, participants had reliable improvements from pretest (*M* = .34, *SD* = .16) to immediate posttest (*M* = .54, *SD* = .20), *t*(74) = 7.55, *p* < .001, *d* = 1.10, and to delayed test (*M* = .47, *SD* = .18), *t*(74) = 5.05, *p* > .001, *d* = .76. The drop between post to delayed test was small but reliable, *t*(74) = 2.93, *p* > .01, *d* = .37. There were no main effect of condition and no phase x condition interaction, *p*'s > .10.

Fluent Accuracy Gain. The 2 phase x 3 condition ANCOVA also showed little forgetting between immediate posttest and delayed test, and no differences in conditions in terms of the amount gained. There were no main effect of phase, *F*(1,71) = .24, *p* = .63, η^2_p = .003, nor of condition, *F*(2, 71) = 1.38, *p* = .26, η^2_p = .04, and no interactions, *p*'s > .10.

The patterns of results were different between Sine and Exponential TF/NI (*Figures D.1b and D.2b)*. There was no condition differences on Sine TF/NI items, *p*'s > .10, but on Exponential TF/NI items, the 3 phase x 3 condition ANOVA confirmed a phase x condition interaction, *F*(4,144) = 3.13, *p* = .02, η^2_p = .08, and no main effect of condition, *p* > .10. At pretest, there were already condition differences on Exponential TF/NI, with both the *passive* group starting out marginally higher and *active* group significantly higher than the *passive-active* group (29%, 30% and 19%, respectively), *t*(48) = 1.79, *p* = .08, *d* = .51, and *t*(48) = 2.27, *p* = .03, *d* = .65, respectively. These differences were of medium effect sizes. At immediate posttest, the *active* and *passive-active* groups in turn did better than the *passive* group (57% and 53% versus 40%, *t*(48) = 2.43, *p* = .02, *d* = .69, and *t*(48) = 1.66, *p* = .10, *d* = .47, respectively). There were no condition differences at delayed test, *p*'s > .10.

Untrained Functions (UF)

Fluent Accuracy. *Figure D.3b* shows the fluent accuracy on UF by condition. Participants from all three groups also improved in fluency on UF items. The 3 phase x 3 condition ANOVA showed a main effect of phase, $F(2,144) = 12.55$, $p < .001$, $\eta^2_p = .15$. The gain from pretest ($M = .29$, $SD = .15$) to immediate posttest ($M = .41$, $SD = .19$) was reliable, $t(74) = 4.42$, $p < .001$, $d = .70$, so was the gain from pretest to delayed test ($M = .41$, $SD = .19$), $t(74) = 4.79$, $p < .001$, $d = .70$, with no loss between post and delayed test, $t(74) = .11$, $p = .92$. Interestingly, there was a marginally significant main effect of condition, $F(2,72) = 2.44$, $p = .09$, $\eta^2_p = .06$. There was no difference between the *active* ($M = .34$, $SD = .10$) and *passive* conditions ($M = .35$, $SD = .12$), on overall UF score, $p > .10$, but the *passive-active* group ($M = .40$, $SD = .08$) produced higher overall UF score than the *active* group, $t(48) = 2.26$, $p = .03$, $d = .66$, and marginally higher than the *passive* group, $t(48) = 1.75$, $p = .09$, $d = .49$. There was no phase x condition interaction, $p > .10$.

Fluent Accuracy Gain. There were no effects of condition nor phase x condition interaction on the fluency gain, p's > .10.

Combination Functions (CF)

Figure D.3b shows the fluent accuracy on CF by condition. There were no significant effects, p's > .10.

Response Times on Correct Answers (RTc)

Across all conditions, participants had very small and marginally significant improvements on RTc, $F(2,144) = 2.51$, $p = .09$, $\eta^2_p = .03$, but pairwise comparisons did not confirm statistically significant differences. Participants needed about 7.51 seconds per correct answer at pretest ($SD = 3.26$), took 8.13 seconds ($SD = 2.84$) at immediate posttest, and 7.39

seconds (SD = 2.99) at delayed test. There were also no main effect of condition nor phase x condition interaction, *p*'s $> .10$.

APPENDIX E

Full List of Assessment Items used in Experiments 2, 4, 6

	Item Type	Sub-category	Version A	Version B	Version C
1	TI	Sine *x*-shift right	y = Sin(x - 3)	y = Sin(x - 2)	y = Sin(x - 4)
2	TI	Sine *x*-scale compression	y = Sin(3x)	y = Sin(4x)	y = Sin(2x)
3	TI	Sine *y*-shift up	y = Sin(x) + 2	y = Sin(x) + 4	y = Sin(x) + 3
4	TI	Sine *y*-scale expansion	y = Sin(x) / 4	y = Sin(x) / 3	y = Sin(x) / 2
5	TI	Exponential *x*-shift left	y = Exp(x + 4)	y = Exp(x + 3)	y = Exp(x + 2)
6	TI	Exponential *x*-scale expansion	y = Exp(x / 2)	y = Exp(x / 4)	y = Exp(x / 3)
7	TI	Exponential *y*-shift down	y = Exp(x) - 3	y = Exp(x) - 2	y = Exp(x) - 4
8	TI	Exponential *y*-scale compression	y = 2 * Exp(x)	y = 3 * Exp(x)	y = 4 * Exp(x)
9	TF/NI	Sine *x*-shift left	y = Sin(x + 30)	y = Sin(x + 40)	y = Sin(x + 20)
10	TF/NI	Sine *y*-scale compression	y = 40 * Sin(x)	y = 20 * Sin(x)	y = 30 * Sin(x)
11	TF/NI	Sine *y*-shift down	y = Sin(x) - 20	y = Sin(x) - 30	y = Sin(x) - 40
12	TF/NI	Sine *x*-scale expansion	y = Sin(x / 30)	y = Sin(x / 40)	y = Sin(x / 20)
13	TF/NI	Exponential *x*-shift right	y = Exp(x - 40)	y = Exp(x - 20)	y = Exp(x - 30)
14	TF/NI	Exponential *y*-scale expansion	y = Exp(x) / 20	y = Exp(x) / 30	y = Exp(x) / 40
15	TF/NI	Exponential *y*-shift up	y = Exp(x) + 30	y = Exp(x) + 40	y = Exp(x) + 20
16	TF/NI	Exponential *x*-scale compression	y = Exp(40x)	y = Exp(20x)	y = Exp(30x)
17	UF	Cos *x*-scale compression	y = Cos(4x)	y = Cos(2x)	y = Cos(3x)
18	UF	Cos *y*-scale expansion	y = Cos(x) / 2	y = Cos(x) / 3	y = Cos(x) / 4
19	UF	Cos *y*-shift up	y = Cos(x) + 3	y = Cos(x) + 4	y = Cos(x) + 2
20	UF	Cos *x*-shift right	y = Cos(x - 4)	y = Cos(x - 2)	y = Cos(x - 3)
21	UF	Log *y*-scale compression	y = 2 * Log(x)	y = 3 * Log(x)	y = 4 * Log(x)
22	UF	Log *x*-scale expansion	y = Log(x / 4)	y = Log(x / 2)	y = Log(x / 3)
23	UF	Los *y*-shift down	y = Log(x) - 2	y = Log(x) - 3	y = Log(x) - 4
24	UF	Log *x*-shift left	y = Log(x + 3)	y = Log(x + 4)	y = Log(x + 2)
25	CF	Exponential *x*-scale compression and *y*-shift up	y = Exp(4x) + 4	y = Exp(3x) + 3	y = Exp(2x) + 2
26	CF	Basic Sine and Exponential *y*-shift up	y = Sin(x) + Exp(x) + 3	y = Sin(x) + Exp(x) + 2	y = Sin(x) + Exp(x) + 4
27	CF	Sine *y*-scale compression and *x*-shift left	y = 2 * Sin(x + 2)	y = 4 * Sin(x + 4)	y = 3 * Sin(x + 3)
28	CF	Basic Sine and basic Exponential	y = Sin(x) + Exp(x)	y = Sin(x) + Exp(x)	y = Sin(x) + Exp(x)

APPENDIX F

F.1. Experiment 3 - Results from All Participants

Since we had unequal number of participants in each condition (N = 122, 90 female), the following analyses were conducted with weighted means. All assumptions were met.

Efficiency

Efficiency by Trials

A 2 phase (pre-post, pre-delayed) x 3 conditions (*single*, *contrastive*, *mixed*) ANCOVA on the efficiency scores calculated by trials invested in the training, with pretest accuracy as the covariate, confirmed a significant phase x condition interaction, $F(2,118) = 3.83$, $p = .025$, $\eta^2_p = .06$. There were no condition differences on pre-post efficiency, but on pre-delayed efficiency, the *single* condition ($M = .002$, $SD = .002$) significantly and the *mixed* condition ($M = .001$, $SD = .001$) marginally outperformed the *contrastive* condition ($M = .001$, $SD = .002$), $t(76) = 2.45$, $p = .02$, $d = .73$, and $t(82) = 1.94$, $p = .06$, $d = .51$. The *mixed* and *single* conditions did not differ on trial efficiencies, p's > .10. There were no other effects, p's > .10.

There was a main effect of pretest, $F(1,118) = 8.87$, $p = .004$, $\eta^2_p = .07$. Pretest accuracy did not correlate with pre-post efficiency by trial, $r(122) = -.13$, $p = .17$, but the higher the pretest, the lower the pre-delayed efficiency, $r(122) = -.44$, $p < .001$.

Efficiency by Time

A 2 phase (pre-post, pre-delayed) x 3 conditions (*single*, *contrastive*, *mixed*) ANCOVA on the efficiency scores calculated by minutes invested in the training, with pretest accuracy as the covariate revealed a marginally significant main effect of phase, $F(1,118) = 3.72$, $p = .06$, $\eta^2_p = .03$. The pre-post time efficiency ($M = .008$, $SD = .006$) was significantly higher than the pre-

delayed efficiency (M = .004, SD = .006). There was no main effect of condition, $p > .10$, but there was a significant phase x condition interaction, $F(2, 118) = 5.29$, $p < .01$, $\eta^2_p = .08$. There were no condition differences on immediate posttest efficiency (p's > .05) but on delayed test efficiency, the *single* condition (M = .007, SD = .007) and the *mixed* condition (M = .005, SD = .005) gave higher delayed efficiency than the *contrastive* condition (M = .002, SD = .005), $t(80) = 1.28$, $p = .002$, d = .80, and $t(82) = 2.45$, $p = .016$, d = .60.

There was also a main effect of pretest on the dependent variables, $F(1,118) = 17.78$, $p < .001$, $\eta^2_p = .13$. Pretest accuracy correlated strongly with pre-post efficiency by time, $r(122) = -.26$, $p < .01$, and to pre-delayed efficiency by time, $r(122) = -.43$, $p < .001$.

Accuracy

A 3 phase (pre, post, delayed) x 3 conditions (*single*, *contrastive*, *mixed*) ANOVA confirmed a main effect of phase, $F(2, 238) = 156.13$, $p < .001$, $\eta^2_p = .57$. In general, regardless of condition, participants improved from pretest (M = .27, SD = .18) to immediate posttest (M = .55, SD = .17), $t(121) = -16.99$, $p < .001$, d = 1.84, and from pretest to delayed test (M = .43, SD = .16), $t(121) = 9.32$, $p < .001$, d = 1.09. The drop from immediate posttest to delayed test was also statistically significant, $t(121) = 8.04$, $p < .001$, d = .73.

There was an interaction of phase x condition, $F(4, 238) = 3.96$, $p < .01$, $\eta^2_p = .06$. There were no condition differences at pretest and at immediate posttest (p's > .10), but at delayed test, the *single* group (M = .48, SD = .15) and the *mixed* group (M = .45, SD = .15) retained much more of what they have learned than the *contrastive* group (M = .37, SD = .18), $t(76) = 2.89$, $p = .005$, d = .66, and $t(82) = 2.22$, $p = .03$, d = .48. There was no main effect of condition, $F(2,119) = .49$, $p = .61$, $\eta^2_p = .008$.

Accuracy Gain

A 2 phase (pre-post, pre-delayed) x 3 conditions (*single*, *contrastive*, *mixed*) ANCOVA with pretest accuracy as the covariate showed a marginal main effect of condition, $F(2, 118) = 2.33$, $p = .10$, $\eta^2_p = .04$. Overall, both the *single* group ($M = .25$, $SD = .19$) and the *mixed* group ($M = .25$, $SD = .15$) produced higher learning gain than the *contrastive* group ($M = .17$, $SD = .17$), $t(76) = 2.08$, $p < .05$, $d = .44$, and $t(80) = 2.15$, $p < .05$, $d = .50$.

There was also a phase x condition interaction, $F(2,118) = 5.03$, $p < .01$, $\eta^2_p = .08$. There were no condition differences in pre-posttest gain, $p > .10$, but in terms of pre-delayed test gain, the *single* group ($M = .22$, $SD = .19$) and the *mixed* group ($M = .19$, $SD = .17$) outperformed the *contrastive* group ($M = .08$, $SD = .20$), $t(76) = 3.24$, $p = .002$, $d = .72$, and $t(82) = 2.61$, $p = .01$, $d = .59$, respectively. There was no difference between the *single* and the *mixed* group, $p > .10$, and no phase x pretest interaction, $F(1,118) = 1.17$, $p > .10$, $\eta^2_p = .01$.

There was a main effect of the pretest, $F(1, 118) = 225.88$, $p < .001$, $\eta^2_p = .34$. Pretest accuracy correlated strongly with pre-post, $r(122) = -.48$, $p < .001$, and pre-delayed learning gain, $r(122) = -.57$, $p < .001$. There was a main effect of phase, $F(1,118) = 7.71$, $p < .01$, $\eta^2_p = .06$. The gain from pre-post test ($M = .28$, $SD = .19$) was significantly higher than the gain from pre-delayed test ($M = .16$, $SD = .19$).

Response times on correct answers

A 3 phase x 3 condition ANOVA confirmed a main effect of phase, $F(2, 238)= 44.11$, $p <.001$, $\eta^2_p = .27$. All three groups improved in the time needed to reach accurate responses from pretest ($M = 12.15$, $SD = 8.24$) and immediate posttest ($M = 8.53$, $SD = 2.80$), $t(121) = 8.18$, $p <.001$, $d = .59$, and to delayed test ($M = 8.41$, $SD = 3.18$), $t(121) = 6.63$, $p < .001$, $d = .60$. The speed gain from pretest to immediate posttest was perfectly preserved at delayed test a week later (post vs. delayed test, $t(121) = 1.05$, $p >.10$. There were no phase x condition interaction

F(4,238) = .09, $p > .10$, $\eta^2_p = .001$, nor main effect of condition, $F(2,119) = .29$, $p > .10$, $\eta^2_p = .005$.

Fluent Accuracy

There was a main effect of phase, $F(2,238) = 210.86$, $p <. 001$, $\eta^2_p = .64$. In general, participants improved significantly from pretest ($M = .18$, $SD = .12$) to immediate posttest ($M = .50$, $SD = .18$), t(121) = 19.47, $p < .001$, $d = 2.09$, and to delayed ($M = .37$, $SD = .16$), t(121) = 12.05, $p <.001$, $d = 1.34$. The difference between post and delayed test was also statistically significant, t(121) = 8.24, $p <.001$, $d = .76$.

There was no main effect of condition, $p > .10$, but there was a phase x condition interaction, $F(4,238) = 3.33$, $p = .01$, $\eta^2_p = .05$. There were no condition differences at pretest and immediate posttest, but at delayed test, the *single* ($M = .41$, $SD = .17$) group was able to remember much of the learned information and apply them to new instances at delayed test than the *contrastive* condition ($M = .33$, $SD = .17$), t(76) = 2.13, $p = .04$, $d = .47$. There were no other reliable differences at delayed test, $p > .10$.

Fluency Accuracy Gain

A 2 phase x 3 condition ANCOVA with pretest score being the covariate showed a main effect of pretest score, $F(1,118) = 23.03$, $p <.001$, $\eta^2_p = .16$. Pretest negatively correlated with pre-post gain, $r(122) = -.30$, $p < .01$, and with pre-delayed gain, $r(122) = -.45$, $p < .001$. There was a main effect of phase, $F(1,118) = 10.84$, $p < .001$, $\eta^2_p = .08$. The gain from pretest to immediate posttest scores ($M = .31$, $SD = .18$) was statistically higher than the gain from pretest to delayed test ($M = .19$, $SD = .17$), t(121) = 8.24, $d = .69$. There was no main effect of condition, $p > .10$, but there was a phase x condition interaction, $F(2, 118) = 3.12$, $p < .05$, $\eta^2_p = .05$. In terms of pre-delayed score gain, *single* ($M = .25$, $SD = .18$) and *mixed* ($M = .21$, $SD = .15$)

did better than *contrastive* (M= .12, *SD* = .17), $t(76) = 3.20$, $p = .002$, $d = .74$, and $t(82) = 2.48$, $p = .05$, $d = .56$, respectively. There were no condition differences at pre-post score, and no phase x pretest interaction, p's > .10.

Progression of Learning

Table F.1 shows the training means for all participants and separately for participants who have reached learning criteria (i.e., completed the modules). There were no discerning condition differences in training accuracies. The *mixed* condition started out with lower accuracy numerically than the other condition, but pairwise comparisons did not confirm any reliable differences across blocks (or quartiles), p's > .05.

	Conditions	**Trials Completed**	**Minutes on Module**	**Training Accuracy**	**Percent Mastery**	**Training Fluency**	**Blocks completed**
All (N = 122)	*Single* (N = 38)	142.4 (9.5)	40.76 (2.46)	.53 (.02)	.87 (.05)		11.86 (.79)
	Contrastive (N = 40)	146.6 (9.4)	44.10 (2.63)	.50 (.02)	.81 (.06)		12.21 (.78)
	Mixed (N = 44)	148.2 (7.9)	43.62 (2.06)	.48 (.02)	.75 (.06)		12.03 (.59)
Completed (N = 90, 30 per condition)	*Single*	136.00 (10.58)	36.83 (2.42)	.57 (.02)	1.00	.52 (.02)	11.33 (.88)
	Contrastive	139.90 (10.42)	36.83 (13.61)	.56 (.02)	1.00	.49 (.02)	11.75 (.76)
	Mixed	140.97 (9.06)	39.60 (2.61)	.54 (.02)	1.00	.49 (.02)	11.66 (.87)
	-Single trials	70.30 (4.69)		.55 (.02)		.51 (.02)	
	-Contrastive trials	70.70 (4.66)		.53 (.02)		.46 (.01)	

Table F.1. Training means of all participants (top half) and of participants who have completed the training modules (bottom half). Standard errors are in parentheses.

Survey

There were no condition differences on participants' self-ratings of how enjoyable the training was, how motivated and engaged participants were, how much they have learned, how much they will remember a week from now, nor on the level of helpfulness of the training module, *p*'s > .05. However, there was a main effect of condition on how helpful participants found the primer to be, $F(2,118) = 4.09$, $p = .02$. This was driven by differences between the *contrastive* and *single* group, and between the *mixed* and *single* group. Both of the *contrastive* ($M = 4.05$, $SD = 1.11$) and *mixed* ($M = 4.23$, $SD = 1.29$) groups rated the primer to be higher in helpfulness than the *single* training group ($M = 3.47$, $SD = 1.29$), $t(76) = 2.12$, $p = .04$, $d = .48$, and $t(79) = 2.65$, $p = .01$, $d = .59$, respectively. The difference between the *mixed* and *single* groups was not statistically significant, $p > .05$.

Dweck Theory of intelligence questions

We averaged the responses over 4 theory of intelligence items, the higher the rating, the more the participant endorsed a fixed mindset. This did not correlate with percent mastery, training accuracy or efficiency. Interestingly, this correlated with the pretest accuracy, suggesting that those with higher fixed mindset ratings performed better on the pretest (but only the pretest), $r(121) = .22$, $p = .01$. As a result, fixed theorists were more likely to have a lower pre-post accuracy gain, $r(121) = -.29$, $p = .002$, and pre-delayed test accuracy $r(121) = -.17$, $p = .06$. Those with higher fixed ratings were also more likely to rate themselves lower on how much they learned $r(121) = -.20$, $p = .025$ and marginally more likely to give lower rating to how helpful the module was, $r(121) = -.16$, $p = .08$.

F.2. Experiment 3 - Extra Analyses and Detailed Results

Of those who started the modules but did not finish, 16 were in the *single* condition, 18 in the *contrastive* condition, and 24 in the *mixed* condition. *Table F.1* shows the mean training performance from the participants who did not finish.

Condition	Learning Accuracy	Minutes Spent	Trials completed	Pretest accuracy	Percent Mastery
Single	.16 (.01)	89.2 (28.8)	390 (123.1)	.15 (.02)	.24
Contrastive	.16 (.01)	109.4 (59.49)	267 (72.9)	.14 (.02)	.09
Mixed	.16 (.01)	58.58 (13.16)	209.5 (39.2)	.19 (.02)	.16

Table F.1. Training means of participants who dropped out during the training.

The following analyses were conducted on data from participants who have reached learning criteria.

Time Efficiency

A 3 phase x 3 condition ANOVA on time efficiency confirmed a main effect of phase, $F(1, 86) = 6.40$, $p = .01$, $\eta^2_p = .07$, such that the pre-post efficiency ($M = .009$, $SD = .006$) was higher than the pre-delayed efficiency ($M = .006$, $SD = .006$). There was also a phase x condition interaction, $F(2, 86) = 5.28$, $p = .007$, $\eta^2_p = .11$. While there was no differences in condition on pre-post efficiency, both the *single* ($M = .008$, $SD = .007$) and *mixed* ($M = .006$, $SD = .005$) conditions produced greater pre-delayed gain than the *contrastive* condition ($M = .003$, $SD = .006$) with medium effect sizes, $t(58) = 3.11$, $p = .003$, $d = .80$; $t(58) = 2.34$, $p = .02$, $d = .60$, respectively. There were no other effects, p's > .10.

The pretest strongly predicted the training efficiency, $F(1, 86) = 26.39$, $p < .001$, $\eta^2_p = .24$, such that the lower the pretest accuracy, the greater the pre-post efficiency, $r(90) = -.40$, $p < .001$, and the pre-delayed efficiency, $r(90) = -.51$, $p < .001$.

Accuracy

The 2 phase (pre-post, pre-delayed) x 3 condition ANCOVA with pretest accuracy as the covariate supported a main effect of phase, $F(1,86) = 11.15$, $p = .001$, $\eta^2_p = .12$, reflecting a reliable drop in memory between the pre-post gain ($M = .31$, $SD = .18$) and the pre-delayed gain ($M = .18$, $SD = .20$). There was also a significant interaction effect of phase x condition, $F(2, 86) = 4.95$, $p = .009$, $\eta^2_p = .103$. There were no condition differences in pre-post gain, however, both the *single* ($M = .25$, $SD = .18$) and *mixed* ($M = .21$, $SD = .16$) conditions produced greater pre-delayed gain than the *contrastive* condition ($M = .09$, $SD = .21$), $t(58) = 3.13$, $p = .003$, $d = .82$; $t(58) = 2.39$, $p = .02$, $d = .64$, respectively. There was no interaction of phase x pretest, $F(1,86) < 1$, $p > .10$, nor main effect of condition, $F(2, 86) = 1.98$, $p > .10$, $\eta^2_p = .04$.

Pretest accuracy reliably predicted posttest gains, $F(1,86) = 81.36$, $p < .001$, $\eta^2_p = .49$. The lower the pretest, the higher the gain at immediate posttest, $r(90) = -.63$, $p < .001$, and at delayed test, $r(90) = -.63$, $p < .001$.

Response Times on Correct Answers

Participants in all three conditions had very similar response times on correct answers, p's > .10. There were no main effect of condition nor a phase x condition interaction, p's > .10, but participants improved dramatically in the time required for correct answers following the training, $F(2, 174) = 51.86$, $p < .001$, $\eta^2_p = .37$. At pretest, they needed an average of 12.16 seconds ($SD = 4.83$) and improved by nearly 46% at immediate posttest ($M = 7.86$, $SD = 2.18$), $t(89) = 9.28$, $p < .001$, $d = 1.15$, and much of this speed gain was maintained at delayed test ($M =$

8.72, *SD* = 3.12), pre vs. delayed test: $t(89) = 6.45$, $p < .001$, $d = .85$. The difference between immediate posttest and delayed test was statistically significant but with a small effect size, $t(89) = 2.90$, $p < .01$, $d = .32$.

Fluent Accuracy

The pattern of learning gains was similar with that shown by accuracy. Both of the *single* and *mixed* conditions produced stronger pre-delayed learning gain, suggesting that *single* trials were important for long-term retention. The *mixed* condition did not differ from the *contrastive* condition at delayed test.

A 3 phase (pre, post, delayed) x 3 conditions ANOVA confirmed these observations. There was a main effect of phase, $F(2, 174) = 213.32$, $p < .001$, $\eta^2_p = .71$. All conditions produced reliable learning gain in fluent accuracy from pretest (*M* = .20, *SD* = .12) to immediate posttest (*M* = .55, *SD* = .15), $t(89) = 21.25$, $p < .001$, $d = 2.58$, and to delayed test (*M* = .41, *SD* = .15), $t(89) = 11.77$, $p < .001$, $d = 1.55$, though there was a reliable drop between post and delayed test as well, $t(89) = 7.74$, $p < .001$, $d = .93$. There was a phase x condition interaction, $F(4,174) = 2.92$, $p = .023$, $\eta^2_p = .06$. At delayed test, the *single* condition performed better (*M* = .45, *SD* = .15) than the *contrastive* condition (*M* = .37, *SD* = .17), $t(58) = 2.08$, $p = .04$, $d = .50$. This difference was not statistically significant between the *mixed* (*M* = .42, *SD* = .12) and the *contrastive* condition, $t(58) = 1.3$, $p = .17$, $d = .34$. There was no difference between *mixed* and *single*, $p > .10$, and no main effect of condition, $p > .10$.

Fluent Accuracy Gain

A 2 phase (pre-post, pre-delayed) x 3 conditions (*single*, *contrastive*, *mixed*) ANCOVA with pretest fluent accuracy as the covariate confirmed a significant main effect of pretest score, $F(1,86) = 37.48$, $p < .001$, $\eta^2_p = .30$, such that the pretest score reliably and negatively predicted

both the pre-post score gain, $r(90) = -.46$, $p < .001$ and the pre-delayed score gain, $r(90) = -.51$, $p < .001$. There was also a significant main effect of phase, $F(1, 86) = 12.54$, $p = .001$, $\eta^2_p = .13$, reflecting the drop between pre-post gain ($M = .36$, $SD = .16$) and the pre-delayed gain ($M = .22$, $SD = .17$). Interestingly, there was a marginally significant interaction of phase x condition, $F(2,86) = 2.96$, $p = .06$, $\eta^2_p = .06$. While the three groups did not differ at immediate posttest gain (p's > .10), in terms of pre-delayed test gain, the *single* training condition ($M = .27$, $SD = .18$) and the *mixed* condition ($M = .24$, $SD = .14$) performed significantly better than the *contrastive* condition ($M = .14$, $SD = .18$), $t(58) = 2.72$, $p = .009$, $d = .72$, and $t(58) = 2.20$, $p = .03$, $d = .62$, respectively. There were no other condition differences, p's > .10, and no other effects, p's > .10.

Progression of Learning

Accuracy

Figures F.1a and F.1b show the average accuracy and fluency, respectively, by training quartiles for each condition. Overall there were no discerning differences in training performance among the three conditions. A 4 quartile x 3 condition ANOVA confirmed these findings. The presence of the Normal ECG half did not enhance accuracy on *contrastive* items than the other two conditions, $p > .10$. The *mixed* condition provided a within-subject comparison of the single versus contrastive trials, and there was no difference in accuracy between the single and the contrastive trials, $p > .10$. All three groups produced consistent and steady increases in accuracy throughout the training, $F(3,261) = 153.58$, $p < .001$, $\eta^2_p = .64$, from 39% ($SD = 12.64$) at the 1st quartile to 69% ($SD = 11.76$) at the 4th quartile, $t(89) > 5.02$, $p < .001$, $d = .51$ to 1.13.

Fluency Accuracy

There were no differences across conditions on the fluent accuracy, $F(2,87) = 1.48$, $p >$ .10.

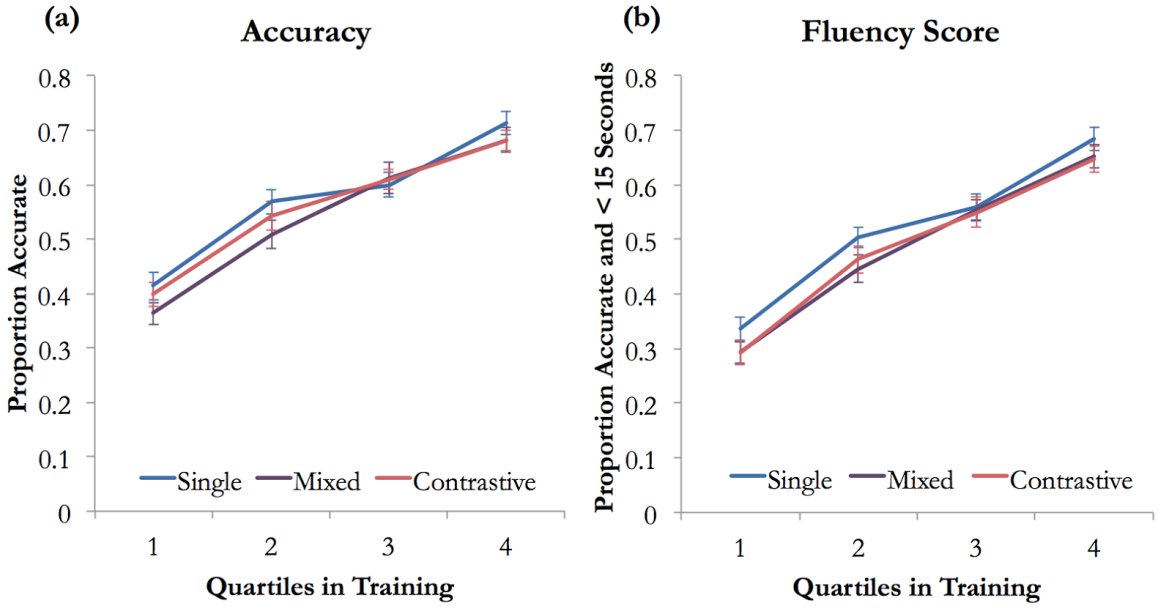


Figure F.1. Average (a) accuracy and (b) fluent accuracy during the training.

Theory of intelligence questions

We averaged the responses over 4 theory of intelligence items. This did not correlate with percent mastery, training accuracy or efficiency. Interestingly, this correlated with the pretest accuracy, suggesting that those with higher fixed mindset ratings performed better on the pretest (but only the pretest), $r(89) = .26$, $p = .01$. As a result, fixed theorists were more likely to have a lower pre-post accuracy gain, $r(89) = -.28$, $p = .008$, and marginally lower pre-delayed test accuracy gain, $r(89) = -.18$, $p = .09$. Those with higher fixed ratings were also more likely to rate themselves lower on how much they learned, $r(89) = -.27$, $p = .01$ and marginally lower on how helpful the module was, $r(89) = -.20$, $p = .06$.

APPENDIX G
Experiments 4 & 6 - Demographic Data

We report demographic data for these experiments together because they were ran at the same time (N = 120).

Age

Age ranged 18-60, with the majority (71%) between 23-35; mean age = 32.01, *SD* = 9.25.

Location

Participants came from 39 US states.

Ethnicity

80% White, 8.3% Asian/Pacific Islander, 4.2% Black or African American, 2.5% Hispanic or Latino/a, 4.2% Mixed, and the rest did not say.

Math background

The majority (81.7%) reported that their last math class was more than 2 years ago, 8.3% 1-2 years ago, 4.2% within a year ago, and 5.8% were taking a math class at the time of the study.

In response to "How much math is involved in your current job on a scale from 1-6 (1 = not at all, 6 = it's all math!)", 72.5% of participants chose 1-3 indicating that their jobs do not require much math. 19.2% chose 4, and only 8.3% chose 4 and 5.

Education

In terms of the highest educational level achieved, 32.5% of participants had some college, 31.7% had a Bachelor's degree, 18.3% had a graduate or professional degree, 8.3% had a high school diploma, 6.7% had trade/technical/vocational training, and .8% (1 person) had some high school but no diploma. The rest (1.7%) did not provide a response.

English Fluency

On a scale from 1-5, with 5 being native/near-native, 97.5% of participants chose 5, and the rest chose 4.

Theory of Intelligence and Math Attitude

We averaged the responses over 4 theory of intelligence questions, and recoded the averaged responses below 3.5 (midpoint) as growth theorists, and above 3.5 as fixed theorists (M = 3.06, SD = 1.55). 42.5% of our participants were fixed theorists, and 57.5% were growth theorists.

Expectedly, those with a growth mindset tended to have a more positive attitude toward math (M = 2.80, SD = .81), $r(120) = -.30$, $p = .001$. Interestingly, however, those with a more fixed mindset were more likely to do better on "Describe the transformations" items at delayed test, $r(120) = .22$, $p = .02$.

APPENDIX H

Experiment 4 Extra Analyses and Detailed Results

Accuracy

Accuracy gain

The ANCOVA with pretest as a covariate confirmed an overall difference in accuracy between conditions. The main effect condition did not reach statistical significance but had a moderate effect size, $F(2,68) = 2.33$, $p = .11$, $\eta^2_p = .06$. The *single* condition showed stronger overall accuracy gain than the *contrastive* condition (19% vs. 12%, respectively, $t(46) = 2.09$, $p = .04$, $d = .60$), and the *mixed* condition also showed marginally stronger gain than the *contrastive* condition (19% vs. 12%, respectively, $t(46) = 1.65$, $p = .10$, $d = .48$.), both with medium effect sizes. At immediate posttest, both *single* ($M = .2$, $SD = .13$) and *mixed* ($M = .21$, $SD = .15$) were marginally better than *contrastive* ($M = .13$, $SD = .15$), $t(46) = 1.88$, $p = .07$, $d = .54$, and $t(46) = 1.84$, $p = .07$, $d = .53$, respectively. At delayed test, the *single* condition did marginally better than the *contrastive* with a medium effect size ($M = .18$, $SD = .11$, and $M = .11$, $SD = .13$, respectively), $t(46) = 1.87$, $p = .07$, $d = .54$. There was no reliable difference between the *mixed* and *contrastive* conditions at delayed test, $p > .10$.

There was also a main effect of the pretest on the accuracy gain, $F(1,68) = 11.50$, $p < .01$, $\eta^2_p = .15$. The higher the pretest, the smaller the learning gain at immediate posttest, $r(72) = -.36$, $p < .001$, and at delayed test, $r(72) = -.33$, $p < .001$.

Trained Items (TI)

Raw accuracy

All conditions produced strong learning, $F(2,138) = 54.32$, $p < .001$, $\eta^2_p = .44$. Across modules, training boosted accuracy on TI from pretest to immediate posttest (16% to 42%, $t(71)$

= 10.24, $p < .001$, $d = 1.57$), and from pretest to delayed test (16% to 34%, $t(71) = 6.77$, $p < .001$, $d = 1.09$). This was true for all conditions from pretest to immediate posttest (s*ingle*: $t(23) = 8.43$, $p < .001$, $d = 2.13$; *contrastive*: $t(23) = 5.08$, $p < .001$, $d = 1.18$; *mixed*: $t(23) = 4.93$, $p < .001$, $d = 1.51$) and from pretest to delayed test (*single*: $t(23) = 5.17$, $p < .001$, $d = 1.44$; *contrastive*: $t(23) = 3.31$, $p < .01$, $d = .80$; *mixed*: $t(23) = 3.50$, $p < .01$, $d = 1.05$). *Table H.1* summarizes these findings for each condition by item types. There was also an overall drop between immediate posttest and delayed test, $t(71) = 3.31$, $p < .01$, $d = .43$.

There was a marginal main effect of condition, $F(2,69) = 2.81$, $p = .07$, $\eta^2_p = .08$, with the *single* condition doing better overall than the *contrastive* condition, $t(46) = 2.36$, $p = .02$, $d = .68$. This was driven by condition differences on Sine TI, $t(46) = 2.38$, $p < .05$, $d = .68$, but not Exponential TI items, $p > .10$.

At pretest, there were no condition differences, p's > .10, but at both post and delayed test, the *single* condition had higher accuracy on TI than the *contrastive* condition (immediate posttest: $t(46) = 2.08$, $p = .04$, $d = .60$, and delayed test, $t(46) = 2.62$, $p = .01$, $d = .76$, with medium effect sizes). The *mixed* condition also had higher accuracy than the *contrastive* condition, but this difference did not reach statistical significance and had a small effect size, $t(46) = 1.61$, $p = .11$, $d = .47$. There were no other condition differences at post and delayed tests, p's > .10.

Accuracy gain

The difference between *single* and *contrastive* conditions was also apparent in terms of accuracy gain, $t(46) = 2.55$, $p < .05$, $d = .74$. This was apparent in terms of pre-post gain, $t(46) = 2.04$, $p = .05$ (marginal), $d = .59$, and pre-delayed gains, $t(46) = 2.23$, $p = .03$, $d = .65$. Similarly,

this was found only on Sine TI, $t(46) = 2.27$, $p < .05$, $d = .65$, and not Exponential TI. There were no other condition differences in accuracy gain, p's > .10.

There was a main effect of pretest, $F(1,68) = 31.35$, $p < .001$, $\eta^2_p = .32$, and no main effect of phase nor any interactions, p's > .10.

		Single			*Contrastive*			*Mixed*		
Item Type	**Phase**	***T(df)***	***p***	***d***	***t***	***p***	***d***	***t***	***p***	***d***
All items	Pre-Post	7.63	***	1.56	4.31	***	0.88	6.79	***	1.39
	Pre-Delayed	7.56	***	1.54	4.07	***	0.83	4.89	***	1.00
TI	Pre-Post	8.43	***	1.72	5.08	***	1.04	4.93	***	1.01
	Pre-Delayed	5.17	***	1.06	3.31	**	0.68	3.50	**	0.71
Sine TI	Pre-Post	4.90	***	1.00	3.00	**	0.61	3.39	**	0.69
	Pre-Delayed	5.94	***	1.21	2.32	*	0.47	1.97	+	0.40
Exponential TI	Pre-Post	6.04	***	1.23	3.63	**	0.74	4.94	***	1.01
	Pre-Delayed	2.67	*	0.54	3.11	**	0.64	3.82	***	0.78
TF/NI	Pre-Post	4.95	***	1.01	3.15	**	0.64	6.94	***	1.42
	Pre-Delayed	4.34	***	0.89	3.99	***	0.81	5.04	***	1.03
Sine TF/NI	Pre-Post	2.29	*	0.47	2.48	*	0.51	4.24	***	0.86
	Pre-Delayed	2.23	*	0.45	3.47	**	0.71	4.51	***	0.92
Exponential TF/NI	Pre-Post	6.09	***	1.24	3.08	*	0.63	6.08	***	1.24
	Pre-Delayed	3.94	***	0.80	2.56	*	0.52	3.29	**	0.67
UF	Pre-Post	3.84	***	0.78	1.62	*ns*		3.18	**	0.65
	Pre-Delayed	4.74	***	0.97	1.59	*ns*		1.90	+	0.39
Cosine	Pre-Post	4.05	***	0.83	1.44	*ns*		3.50	**	0.71
	Pre-Delayed	5.79	***	1.18	1.19	*ns*		2.51	*	0.51
Logarithmic	Pre-Post	1.93	+	0.39	0.89	*ns*		1.44	*ns*	
	Pre-Delayed	1.62	*ns*		0.78	*ns*		0.20	*ns*	
CF	Pre-Post	0.38	*ns*		0.34	*ns*		1.99	+	0.41
	Pre-Delayed	1.43	*ns*		1.25	*ns*		2.50	*	0.51

Table H.1. Summary of accuracy gains by assessment item type. Degrees of freedom (*df*) = 23 for all comparisons. P-value: *** denotes p < .001, ** denotes $p < .01$, * denotes $p < .05$, + denotes $p < .10$, and *ns* means not significant, $p > .10$.

Trained Functions, Novel Items (TF/NI)

Raw accuracy

Participants across conditions also improved on the TF/NI after the training and were able to retain what they learned a week later, $F(2,138) = 48.04$, $p < .001$, $\eta^2_p = .41$. The differences between pretest and each of the posttests were statistically significant, $t(71) = 8.29$, $p < .001$, $d = 1.33$, and $t(71) = 7.78$, $p < .001$, $d = 1.16$, and the drop between post and delayed test was not significant, $t(71) = 1.29$, $p > .10$. This was true for all conditions. All conditions improved from pretest to immediate posttest (*single*: $t(23) = 4.95$, $p < .001$, $d = 1.49$; *contrastive*: $t(23) = 3.15$, $p < .01$, $d = 1.93$; *mixed*: $t(23) = 6.94$, $p < .001$, $d = 1.57$) and from pretest to delayed test (*single*: $t(23) = 4.34$, $p < .001$, $d = 1.21$; *contrastive*: $t(23) = 3.99$, $p < .01$, $d = 1.01$; *mixed*: $t(23) = 5.04$, $p < .001$, $d = 1.22$).

There were no main effect of condition and no phase x condition interaction, p's > .10. Planned comparisons showed that at immediate posttest, the *mixed* condition ($M = .48$, $SD = .19$) did marginally better than the *contrastive* condition ($M = .38$, $SD = .19$), $t(46) = 1.91$, $p = .06$, $d = .47$, but not at delayed test, $p > .10$. There were no differences between the *single* and *contrastive* conditions and between the *single* and *mixed* conditions, p's > .10.

Accuracy gain

There were no main effects nor interaction, except for the main effect of pretest, $F(1,68) = 1.63$, $p < .001$, $\eta^2_p = .40$, such that the pretest was strongly and negatively correlated with the pre-post accuracy gain, $r(72) = -.60$, $p < .001$, and pre-delayed accuracy gain, $r(72) = -.53$, $p < .001$.

Exponential TF/NI

Interestingly, there were condition differences on Exponential TF/NI. In terms of raw accuracy, at immediate posttest, both the *mixed* and *single* conditions outperformed the *contrastive* condition (*single* versus *contrastive*, $t(46) = 2.62$, $p = .01$, $d = .76$, and *mixed* versus *contrastive*, $t(46) = 2.29$, $p = .03$, $d = .66$. However, there were no condition differences at delayed test. In terms of overall accuracy gain, the *single* condition did marginally better than the *contrastive* condition on Exponential TF/NI gain, $t(46) = 1.97$, $p = .06$, $d = .57$.

Sine TF/NI

The *mixed* condition did better numerically than the *single* and *contrastive* conditions at delayed test (50% vs. 40% and 40%, respectively), but the differences were not reliable (p's = .12).

Untrained Functions (UF)

Raw accuracy

There was no condition differences nor interaction, p's > .10. Even though participants were trained on Sine and Exponentials, they were able to improve on their ability to recognize the transformation on Cosine and Logarithmic functions from pretest to immediate posttest, $t(71) = 4.96$, $p < .001$, $d = .78$, and delayed test, $t(71) = 4.53$, $p < .001$, $d = .71$, and remarkably, with great retention after a week, $t(71) = 1.37$, $p > .10$.

Interestingly, the *single* and *mixed* conditions showed improvements from pretest to immediate posttest, but not the *contrastive* condition (*single*: $t(23) = 3.84$, $p < .01$, $d = .99$; *contrastive*: $p > .10$; *mixed*: $t(23) = 3.18$, $p < .01$, $d = .95$) and from pretest to delayed test (*single*: $t(23) = 4.74$, $p < .001$, $d = 1.09$; *contrastive*: $p > .10$; *mixed*: $t(23) = 1.90$, $p = .07$, $d = .55$).

Accuracy gain

The *single* condition exhibited marginally higher learning gain than the *contrastive* condition with medium effect size, $t(46) = 1.98$, $p = .05$, $d = .52$. There was also a main effect of pretest, $F(1,68) = 27.40$, $p < .001$, $\eta^2_p = .29$, and no other significant effects, p's > .10.

Combination Functions (CF)

Raw accuracy

All conditions showed improvements on combination functions, $F(2,138) = 2.93$, $p < .03$, $\eta^2_p = .05$. Interestingly, there was no reliable difference between pretest and immediate posttest, $t(72) < 1$, $p > .10$, but the gain between pretest and delayed test was reliable, $t(71) = 2.95$, $p < .01$, $d = .43$, and marginally so for the gain between immediate posttest and delayed test, $t(71) = 1.96$, $p = .05$, $d = .35$. There were no main effect of condition nor phase x condition interaction, p's > .10.

Interesting, the *mixed* condition was the only one with improvements from pretest to immediate posttest ($t(23) = 1.99$, $p = .06$, $d = .39$) and from pretest to delayed test ($t(23) = 2.50$, $p = .02$, $d = .67$).

Accuracy gain. There were no significant effects, p's > .10.

Fluent Accuracy

Analyses with fluent accuracy showed the same patterns of results as accuracy. All three conditions showed strong overall improvements, particularly on TI and TF/NI items. The *mixed* and *single* conditions improved on UF, but not the *contrastive* condition. The *mixed* condition was the only one with improvements on Combination items.

All items

Fluent Accuracy

Figure 4.3b (in main text) displays the average fluent accuracy on all items. Similar to the accuracy measure, the 3 phase x 3 condition ANOVA showed a main effect of phase, $F(2,138) = 78.54$, $p < .001$, $\eta^2_p = .53$. Participants from all conditions showed strong fluency gain between pretest to immediate posttest (21% - 39%), $t(71) = 10.53$, $p < .001$, $d = 1.51$, and to delayed test (21% - 36%), $t(71) = 9.18$, $p < .001$, $d = 1.28$. There was a small drop between immediate posttest and delayed test, $t(71) = 2.34$, $p < .05$, $d = .21$. There was no main effect of condition nor phase x condition interaction on scores, p's > .10.

Fluent Accuracy Gain

The 2 phase (pre-post, pre-delayed) x 3 condition ANCOVA with pretest fluent accuracy as the covariate showed a marginally significant main effect of condition, $F(2,68) = 2.37$, $p = .10$, $\eta^2_p = .07$. The *single* condition did better than the *contrastive* condition with a medium effect size, $t(46) = 2.14$, $p = .04$, $d = .62$. The *mixed* condition also did marginally better than the *contrastive* condition, but the difference was of a small effect size, $t(46) = 1.66$, $p = .10$, $d = .48$. There were no other condition differences, p's > .10.

TI

Raw scores

Figure 4.4b displays the average fluent accuracy on TIs. There was a main effect of phase, $F(2,138) = 56.13$, $p < .001$, $\eta^2_p = .45$. The differences between pretest and immediate posttest, $t(71) = 10.25$, $p < .001$, $d = 1.61$ and between pretest and delayed test, $t(71) = 7.82$, $p < .001$, $d = 1.21$, had large to very large effect sizes. The difference between post and delayed test was also reliable, $t(71) = 2.82$, $p < .01$, $d = .39$. The main effect of condition was shy from significance, $F(2,69) = 2.29$, $p = .11$, $\eta^2_p = .06$. The *single* condition, however, did reliably

perform better than the *contrastive* condition with a medium effect size, $t(46) = 2.16$, $p < .05$, $d = .62$. There was no other effects, p's > .10.

Fluent Accuracy Gain

There was a marginal main effect of condition, $F(2,68) = 2.59$, $p = .08$, $\eta^2_p = .07$, with the *single* condition performing marginally better overall than the *contrastive* condition, $t(46) = 1.88$, $p = .07$, $d = .54$. There was also a marginally significant main effect of phase, reflecting a drop in fluent accuracy that did not appear with accuracy gain, $F(1, 68) = 3.06$, $p = .09$, $\eta^2_p = .04$. There was a main effect of pretest, $F(1,68) = 23.27$, $p < .001$, $\eta^2_p = .26$.

TF/NI

Raw scores

Figure 4.5b displays the average fluent accuracy on TF/NI items. There was a main effect of phase, $F(2,138) = 55.02$, $p < .001$, $\eta^2_p = .44$, with participants improving a large amount from pretest to immediate posttest, $t(71) = 9.56$, $p < .001$, $d = 1.44$, from pretest to delayed test, $t(71) = 7.93$, $p < .001$, $d = 1.24$, and no forgetting between post and delayed test, $t(71) = 1.28$, $p > .10$. There was no other significant effects, p's > .10. The same patterns were seen with Sine and Exponential TF/NI items.

There was a marginally significant main effect of condition for Sine TF/NI, $F(2,68) = 2.69$, $p = .08$, $\eta^2_p = .07$, but no significant pairwise differences were found, p's > .10. There were no condition differences on Exponential TF/NI items, p's > .10.

Fluent Accuracy Gain

There were no significant effects, p's > .10.

UF

Raw scores

Figure 4.7b displays the average fluent accuracy on UF items. Similarly, there was a main effect of phase, $F(2,138) = 55.02$, $p < .001$, $\eta^2_p = .44$, with large and very large effect sizes for improvements between pretest and immediate posttest, $t(71) = 9.56$, $p < .001$, $d = 1.44$, and pretest and delayed test, $t(71) = 7.93$, $p < .001$, $d = 1.24$, with no forgetting between immediate posttest and delayed test, $t(71) = 1.28$, $p > .10$. There were no main effect of condition and no phase x condition interaction, p's > .10.

Fluent Accuracy Gain. There were no significant effects, p's > .10.

Combination Items (CF)

Raw scores

Figure 4.8b displays the average fluent accuracy on CF. There was a main effect of phase, $F(2,138) = 3.92$, $p < .05$, $\eta^2_p = .05$. The pretest did not differ reliably from the immediate posttest, $p > .10$, but it did differ from the delayed test, $t(71) = 2.32$, $p < .05$, $d = .35$. The gain from immediate posttest and delayed test was also statistically reliable, though with a small effect size, $t(71) = 2.64$, $p < .05$, $d = .38$. There were no other significant effects, p's > .10.

Fluent Accuracy Gain. There were no significant effects, p's > .10.

Progression of Learning

Accuracy by Quartiles

Figure H.1a shows the mean accuracy by training quartiles. All three groups showed steady and reliable improvements from one quartile to the next, $F(3,207) = 102.08$, $p < .001$, $\eta^2_p = 60$. The learning gain was modest and marginally significant between the 1st and 2nd quartiles ($M = .24$, $SD = .10$ to $M = .26$, $SD = .13$), $t(71) = 2.01$, $p = .05$, $d = .21$, but there were greater

and statistically significant improvements in later quartiles (M = .34, SD = .14 in the 3rd quartile and M = .46, SD = .12 in the 4th quartile), $t(71) = 5.42$, $p < .001$, d = .57 and $t(71) = 8.56$, $p < .001$, d = .93, respectively.

In terms of condition differences, there were no main effect of condition nor quartile x condition interaction, p's > .10. At the 1st quartile, the *contrastive* group had slightly higher accuracy than *single* (M = .26, SD = .13 vs. M = .21, SD = .09, $t(46) = 1.68$, p = .10, d = .45), but there were no differences in later quartiles, nor between the other conditions, p's > .10. There were no other differences among conditions, p's > .10, nor among the *contrastive* and *single* trials within the *mixed* condition, p's > .10.

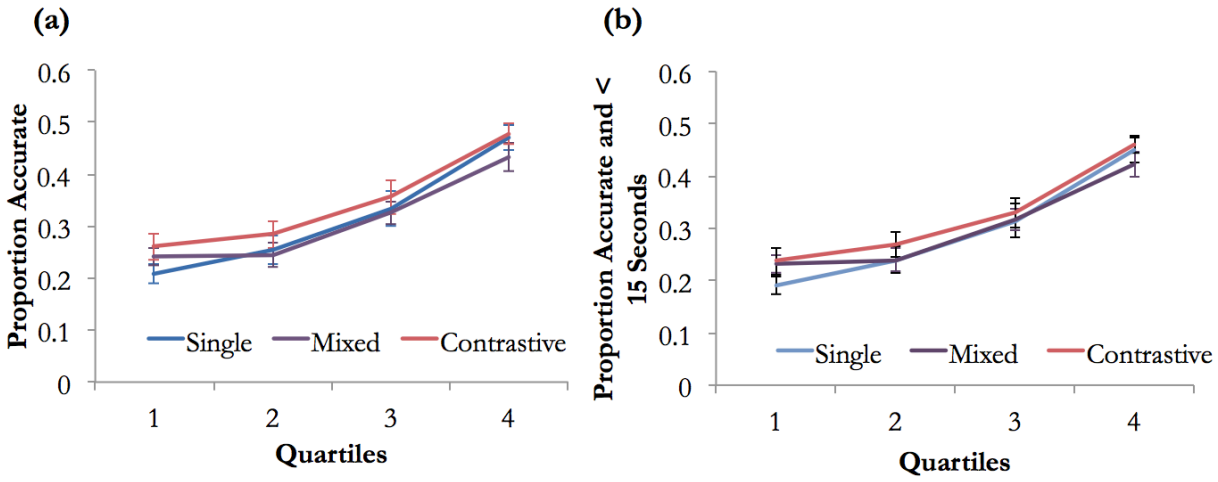


Figure H.1. Training (a) accuracy (b) fluent accuracy by quartiles.

Response Times on Correct Answers (RTc) by Quartiles

Figure 4.9 shows the mean RTc by training quartiles All three groups also showed steady improvements in the RTc across training quartiles, $F(3,207) = 22.06$, $p < .001$, $\eta^2_p = .24$. Participants started out requiring about 7.32 seconds to get reach question correctly (SD = .34) in the 1st quartile, and only needed 5.14 seconds (SD = .28) on the 4th quartile. The gain between

the 1st and 2nd quartiles was reliable with a small effect size, $t(71) = 3.68$, $p < .001$, $d = .31$, and so was the gain between 3rd and 4th quartiles, $t(71) = 4.48$, $p < .001$, $d = .39$, but not between the 2nd and 3rd quartiles, $p > .10$.

Interestingly, the *contrastive* group took longer than the *mixed* group (on contrastive trials) initially (1st quartile: $M = 8.40$, $SD = 2.86$ for *contrastive* vs. $M = 6.67$, $SD = 2.92$ for *mixed*), $t(46) = 2.07$, $p = .04$, $d = .60$; they caught up in later quartiles. The *single* group, on the other hand, started with the same RTc but ended up taking marginally longer in the 4th quartile than the *mixed* group (on single trials only, $M = 4.44$, $SD = 2.00$ for *single*, and $M = 5.75$, $SD = 2.60$ for *mixed*), $t(46) = 1.95$, $p = .06$, $d = .56$. The *mixed* practice may encouraged participants to not dwell as long on contrastive trials, and may play a role in enhancing speed of processing on single trials. The *contrastive* and *single* conditions did not differ on RTc on any quartiles, p's > .10.

Fluent Accuracy by Quartiles

Figure H.1b shows the mean accuracy by training quartiles. At the 1st quartile, *mixed* ($M = .23$, $SD = .08$) did marginally better than *single* ($M = .19$, $SD = .08$), $t(46) = 1.79$, $p = .08$, $d = .53$, but there were no differences at other quartiles, and no other condition differences in fluent accuracy by quartiles, p's > .10.

APPENDIX I

I.1. Experiment 5 - Results from All Participants

Since we had unequal number of participants in each condition (N = 122, 90 female), the following analyses were conducted with weighted means. All assumptions were met.

Efficiency

Efficiency By trials

A 2 phase (pre-post, pre-delayed) x condition on trial efficiency with pretest accuracy as the covariate showed a significant main effect of pretest, $F(1,104) = 8.26$, $p < .01$, $\eta^2_p = .07$. The pretest accuracy negatively correlated with both the pre-post trial efficiency, $r(108) = -.27$, $p < .01$, and pre-delayed trial efficiency, $r(108) = -.24$, $p = .01$. The lower the performance at pretest, the higher the training efficiency.

There was a significant main effect of phase, $F(1,104) = 14.319$, $p < .001$, $\eta^2_p = .12$. The pre-post efficiency ($M = .002$, $SD = .002$) was reliably higher than the pre-delayed efficiency ($M = .001$, $SD = .001$).

There was a marginally significant phase x condition interaction, $F(2,104) = 2.96$, $p = .056$, $\eta^2_p = .05$. The *AL/AC* condition ($M = .003$, $SD = .002$) was marginally more efficient than the *AL/NC* condition ($M = .002$, $SD = .001$) at immediate posttest, $t(77) = 1.92$, $p = .06$, $d = .43$. There were no other differences across conditions in pre-post and pre-delayed efficiencies, p's > .10. There were no significant main effect of condition and no phase x pretest accuracy interaction, p's > .10.

Efficiency By time

The 2 x 3 condition ANCOVA on time efficiency with pretest accuracy as the covariate confirmed a significant main effect of pretest, $F(1, 104) = 6.48$, $p < .05$, $\eta^2_p = .06$. Pretest

accuracy was negatively correlated with the pre-post time efficiency, $r(108) = -.24$, $p < .05$, and with pre-delayed time efficiency, $r(108) = -.22$, $p < .05$.

There was a significant main effect of phase, $F(1, 104) = 15.66$, $p < .001$, $\eta^2_p = .13$. The pre-post time efficiency ($M = .008$, $SD = .006$) was significantly greater than the pre-delayed time efficiency ($M = .005$, $SD = .005$).

There was no significant main effect of condition and no phase x pretest interaction, p's > .10, but there was a significant phase x condition interaction, $F(2, 104) = 3.18$, $p < .05$, $\eta^2_p = .06$. The *AL/AC* condition ($M = .009$, $SD = .006$) proved to be more efficient than the *AL/NC* condition ($M = .006$, $SD = .004$) at immediate posttest, $t(77) = 2.16$, $p = .03$, $d = .48$, and marginally so at delayed test (*AL/AC* $M = .006$, $SD = .005$, *AL/NC* $M = .004$, $SD = .004$), $t(77) = 1.73$, $p = .09$, $d = .39$. The *AL* condition ($M = .009$, $SD = .006$) was also more efficient than the *AL/NC* condition at immediate posttest, $t(71) = 2.10$, $p = .039$, $d = .49$, but not at delayed test, $p > .05$. There were no other differences across conditions at immediate posttest and delayed test, p's > .05.

Accuracy

A 3 phase (pretest, posttest, delayed test) x 3 conditions repeated-measures ANOVA supported a significant main effect of phase, $F(2, 210) = 180.46$, $p < .001$, $\eta^2_p = .63$, and no significant main effect of condition, $F(2,105) = .89$, $p = .41$, $\eta^2_p = .02$, nor a phase x condition interaction, $F(4, 210) = 1.3$, $p = .27$, $\eta^2_p = .02$.

Accuracy gain

A 2 phase (pre-post, pre-delayed) x 3 condition (*AL*, *AL/AC*, *AL/NC*) repeated-measures ANCOVAs with pretest accuracy as the covariate confirmed the same results. There was a main effect of phase, $F(1, 104) = 58.71$, $p < .001$, $\eta^2_p = .36$. Regardless of condition, all participants

were able to transfer what they have learned to new instances at posttests, 26% at pretest to 60% at posttest, t(107) = 17.79, $p < .001$, $d = 2.14$, and at delayed test (49%), t(107) = 11.66, $p < .001$, $d = 1.34$. They also forgot a significant amount after a week of no practice, $t(107) = 7.57$, $p < .001$, $d = .65$. There was a significant main effect of pretest, $F(1, 104) = 42.94$, $p < .001$, $\eta^2_p = .29$, suggesting that the covariate was significantly related to the posttest gains. Indeed, the better participants did at pretest, the less learning gain was found at immediate posttest ($r(108) = -.53$, $p < .001$) and delayed test ($r(108) = -.45$, $p < .001$). There were no statistically significant main effect of condition, $F(2, 104) = 1.02$, $p = .37$, $\eta^2_p = .02$, nor phase x pretest interaction, $F(1, 104) = 1.29$, $p = .27$, $\eta^2_p = .01$, and no phase x condition interaction, $F(2, 104) = 2.27$, $p = .11$, $\eta^2_p = .04$.

Response Times on Correct Answers

There were no differences among groups on RTc, but all three groups showed strong improvements on the time needed to arrive at the correct answers. The speed gain at posttest was sustained over a one week delayed for all participants.

A 3 phase (pretest, posttest, delayed-test) x 3 conditions (*AL*, *AL/AC*, *AL/NC*) ANOVA on RTc confirmed these observations. There was a statistically significant main effect of phase, $F(2, 210) = 36.16$, $p < .001$, $\eta^2_p = .25$, and no other significant effects (p's > .10). Participants improved from an average of 12.30 seconds per correct classification at pretest ($SD = 4.59$) to only 8.88 seconds at immediate posttest ($SD = 2.83$), $t(107) = 7.36$, $p < .001$, $d = .90$, to 9.34 seconds at delayed test ($SD = 2.75$), $t(107) = 6.23$, $p < .001$, $d = .78$. The difference between immediate posttest and delayed test RTc was not statistically significant, $t(107) = 1.48$, $p > .10$.

Fluent Accuracy

Similar to accuracy, when we consider only correct answers within 15 seconds, the same pattern applies: The *AL* group produced similar immediate posttest learning gain as the comparison groups, but showed a dramatic drop at delayed test.

A 3 phase (pretest, posttest, delayed-test) x 3 conditions (*AL*, *AL/AC*, *AL/NC*) ANOVA on fluent accuracy confirmed the patterns. There was a main effect of phase, $F(2, 210) = 214.57$, $p < .001$, $\eta^2_p = .67$. Overall, participants from all conditions improved drastically in fluency from pretest ($M = .17$, $SD = .11$) to immediate posttest ($M = .53$, $SD = .18$), $t(107) = 19.54$, $p < .001$, $d = 1.47$, and to delayed test ($M = .42$, $SD = .18$), $t(107) = 14.57$, $p < .001$, $d = 1.69$. Their scores also dropped reliably from immediate posttest to delayed test, $t(107) = 6.55$, $p < .001$, $d = .65$. There were no significant main effect of condition and phase x condition interaction, p's > .10.

Fluent Accuracy Gain

A 2 phase (pre-posttest, pre-delayed test) x 3 condition (*AL*, *AL/AC*, *AL/NC*) repeated-measures ANCOVAs with pretest fluent accuracy as the covariate confirmed the same pattern. There was a main effect of pretest fluent accuracy, $F(1, 104) = 24.43$, $p < .001$, $\eta^2_p = .19$. Pretest score was highly and negatively correlated with the pre-post gain, $r(108) = -.43$, $p < .001$, and with the pre-delayed test gain, $r(108) = -.32$, $p = .001$, suggesting that the lower scorers at pretest benefited the most at posttests.

There was a main effect of phase, $F(1, 104) = 25.06$, $p < .001$, $\eta^2_p = .19$. The pre-posttest gain ($M = .36$, $SD = .19$) was significantly higher than the pre-delayed test gain ($M = .25$, $SD = .18$), $t(107) = 6.51$, $p < .001$, $d = .61$. There was no significant main effect of condition and no phase x pretest interaction, p's > .10.

Unlike the results from completed participants, there was a marginally significant phase x condition interaction, $F(2, 104) = 2.41$, $p = .09$, $\eta^2_p = .04$. There were no reliable condition

differences in pre-post fluency gain, but when the *AL/AC* condition (M = .28, SD = .17) showed marginally stronger pre-delayed fluency gain than the *AL* condition (M = .20, SD = .19), $t(62)$ = 1.80, p = .08, d = .45. There were no other significant differences between condition, p's > .05.

Progression of Learning

Training Trials

The *AL/AC* and *AL/NC* groups received similar total number of trials in the training as the *AL* group (*Table I.1*). A more careful look suggests that although comparison practice did not reduce the total number of training trials, they may have reduced the number of classification trials needed for the noticeable accuracy gains (*Table I.1*). This was confirmed by an analysis of variance (ANOVA) on the number of classification trials for all participants, $F(2, 105) = 3.57$, $p < .05$. The *AL* condition experienced more active classification trials than those in the *AL/AC* condition (158.00 vs. 127.80 trials), $t(62) = 2.36$, $p = .02$, $d = .59$, and than those in the *AL/NC* condition (158.00 vs. 134.32 trials), $t(62) = 2.34$, $p = .02$, $d = .55$. There was no reliable difference between the *AL/AC* and *AL/NC* conditions on the number of active trials seen, $p > .05$.

However, the three groups differed on the amount of time spent on the module, $F(2, 105) = 4.97$, $p < .01$. The *AL/NC* condition took longer on average than the *AL* condition (54.68 vs. 43.59 minutes), t(71) = 3.24, $p = .002$, $d = .79$. There were no other condition differences in time spent on module, $p > .10$.

Groups did not differ on AL accuracy nor AL RTc, p's > .10. However, one-way ANOVA showed a marginal main effect of condition on AL fluent accuracy, $F(2,69) = 2.92$, $p = .06$. *AL/AC* had higher and *AL/NC* had marginally higher overall score than the *AL* condition (.53 and .52 vs. .48, respectively, $t(46) = 2.46$, $p = .02$, $d = .71$, and $t(46) = 1.87$, $p = .07$, $d = .54$. *AL/AC* and *AL/NC* did not differ, $p > .10$.

Condition	Categories Retired (out of 7)	Minutes on Module	AL trials	AA trials	AB trials	AL accuracy	AA accuracy	AB accuracy
All participants (N = 108)								
AL	6.07	43.59	158.00		--	.51 (.02)		

(N = 29)	(.42)	(2.26)	(8.30)					
AL/AC	5.20	49.11	127.80	35.57	14.77	.48 (.02)	.71 (.02)	.48 (.04)
(N = 35)	(.48)	(2.67)	(9.41)	(5.64)	(2.23)			
AL/NC	4.75	54.68	134.32	37.52	17.02	.50 (.02)	.70 (.02)	.58 (.03)
(N = 44)	(.42)	(2.35)	(6.13)	(3.35)	(1.58)			
Completed participants (N = 72)								
AL	7	40.67	154.46	--	--	.53 (.02)		
(N = 24)		(2.72)	(8.73)					
AL/AC	7	42.50	119.54	25.71	11.08	.58 (.02)	.71 (.03)	.52 (.04)
(N = 24)		(2.80)	(6.86)	(3.24)	(1.31)			
AL/NC	7	47.33	131.42	27.79	13.58	.57 (.02)	.72 (.03)	.62 (.05)
(N = 24)		(2.91)	(7.16)	(3.22)	(1.96)			

Table I.1. Training means of all participants (top half) and of participants who have completed the training modules (bottom half). Standard errors are in parentheses.

I.2. EXPERIMENT 5 - EXTRA ANALYSES

Response Times on Correct Answers

There were no differences among groups on RTc, *p*'s >. 10, but all three groups showed strong improvements on the time needed to arrive at the correct answers. The speed gain at posttest was sustained over a one week delay. A 3 phase (pretest, posttest, delayed test) x 3 conditions (*AL*, *AL/AC*, *AL/NC*) ANOVA on RTc confirmed these observations. There was a statistically significant main effect of phase, $F(2, 138) = 36.17$, $p < .001$, $\eta^2_p = .34$, and no other significant effects ($p > .10$). Participants improved from an average of 12.53 seconds per correct classification at pretest (*SD* = 4.51) to only 8.71 seconds at immediate posttest (*SD* = 2.24), $t(71) = 7.35$, $p < .001$, $d = 1.07$, to 9.52 seconds at delayed test (*SD* = 2.61), $t(71) = 5.55$, $p < .001$, $d = .81$. The small increase from immediate posttest to delayed test was statistically significant but with a small effect size, $t(71) = 2.52$, $p = .01$, $d = .33$.

Fluent Accuracy

A 3 phase (pretest, posttest, delayed test) x 3 conditions (*AL*, *AL/AC*, *AL/NC*) ANOVA on fluent accuracy confirmed a significant main effect of phase, $F(2, 138) = 197.94$, $p < .001$, $\eta^2_p = .74$. Overall, participants improved dramatically in fluent accuracy from pretest (*M* = .19, *SD* = .01) to immediate posttest, (*M* = .59, *SD* = .01), $t(71) = 20.27$, $p < .001$, $d = 3.25$, and to delayed test (*M* = .46, *SD* = .02), $t(71) = 13.12$, $p < .001$, $d = 1.88$. The drop from immediate posttest to delayed test was also reliable, $t(72) = 5.90$, $p < .001$, $d = .86$.

There was a marginal main effect of condition on overall fluent accuracy, $F(2, 69) = 3.04$, $p = .06$, $\eta^2_p = .08$. The *AL/AC* condition produced higher overall fluent accuracy (*M* = .44, *SD* = .08) than the *AL* condition (*M* = .38, *SD* = .11), $t(46) = 2.32$, $p = .03$, $d = .67$. The *AL/NC* condition (*M* = .42, *SD* = .09) also scored higher than the *AL*, but the difference was not

statistically reliable, $p > .10$. There was no difference between the two comparison conditions, $p > .10$.

Unlike accuracy, there was also a marginal interaction of phase x condition, $F(4, 138) = 1.98$, $p = .10$, $\eta^2_p = .05$. At pretest and immediate posttest, there were no differences across conditions (p's > .10). However, and most notably, at delayed test, the *AL/AC* condition ($M = .53$, $SD = .16$) produced higher and the *AL/NC* condition ($M = .48$, $SD = .15$) marginally higher fluent accuracy than the *AL* condition with large and medium effect sizes ($M = .38$, $SD = .18$), $t(46) = 2.99$, $p = .004$, $d = .86$, and $t(46) = 1.97$, $p = .06$, $d = .57$, respectively. The *AL*/*AC* and *AL*/*NC* groups did not differ on their delayed test fluent accuracy, $p > .10$.

Fluent Accuracy Gain

A 2 phase (pre-post, pre-delayed-post) x 3 condition (*AL*, *AL/AC*, *AL/NC*) repeated-measures ANCOVAs with pretest fluent accuracy as the covariate confirmed the findings above. There was a significant main effect of pretest fluent accuracy, $F(1, 68) = 44.52$, $p < .001$, $\eta^2_p = .40$. The lower the pretest score, the higher the pre-post fluency gain, $r(72) = -.63$, $p < .001$, and the higher the pre-delayed fluency gain, $r(72) = -.39$, $p = .001$.

After controlling for the effect of the pretest, there was a significant main effect of phase, $F(1, 68) = 20.05$, $p < .001$, $\eta^2_p = .23$. The pre-post fluency gain ($M = .41$, $SD = .17$) was statistically significantly greater than the gain from pretest to delayed test gain ($M = .28$, $SD = .18$), suggesting overall forgetting after a week. There was also a significant main effect of condition, $F(2, 68) = 3.38$, $p = .04$, $\eta^2_p = .09$. This was driven by overall difference between the *AL/AC* ($M = .44$, $SD = .08$) and the *AL* condition ($M = .38$, $SD = .11$), $t(46) = 2.32$, $p = .03$, $d = .67$. The other conditions did not differ reliably overall, $p > .10$. There were no phase x pretest and phase x condition interactions, p's > .10.

APPENDIX J

J.1. Extra Analyses and Detailed Results for Experiment 6

Accuracy

Table J.1 shows the summary of learning gain results on each item type within each condition. For each item type, we showed the pairwise dependent t-test statistics to assess learning gains between pretest and immediate posttest, and pretest to delayed test. Interestingly, all conditions showed strong learning gains on TI, TF/NI, and even UF. However, *AL/AC* and *AL/NC* produced learning gains on Logarithmic UF and on CF at delayed test, but *AL* did not.

		AL			***AL/AC***			***AL/NC***		
Item Type	Phase	*t*	*p*	*d*	*t*	*p*	*d*	*t*	*p*	*d*
All items	Pre-Post	7.63	***	1.56	11.90	***	2.43	9.57	***	1.95
	Pre-Delayed	7.60	***	1.55	10.87	***	2.22	5.86	***	1.20
TI	Pre-Post	8.14	***	1.66	9.64	***	1.97	7.34	***	1.50
	Pre-Delayed	5.06	***	1.03	6.31	***	1.29	4.83	***	0.99
Sine TI	Pre-Post	4.61	***	0.94	7.45	***	1.52	4.43	***	0.90
	Pre-Delayed	5.35	***	1.09	5.01	***	1.02	2.48	*	0.51
Exponential TI	Pre-Post	5.27	***	1.08	6.79	***	1.39	5.89	***	1.20
	Pre-Delayed	2.64	*	0.54	3.65	**	0.75	4.29	***	0.88
TF/NI	Pre-Post	5.38	***	1.10	9.34	***	1.91	7.70	***	1.57
	Pre-Delayed	4.60	***	0.94	5.24	***	1.07	3.56	**	0.73
Sine TF/NI	Pre-Post	2.41	*	0.49	5.56	***	1.13	3.30	**	0.67
	Pre-Delayed	2.36	*	0.48	3.29	**	0.67	1.62	*ns*	0.33
Exponential TF/NI	Pre-Post	6.08	*	1.24	9.93	***	2.03	6.22	***	1.27
	Pre-Delayed	3.92	***	0.80	5.56	***	1.13	3.29	**	0.67
UF	Pre-Post	3.84	***	0.78	5.38	***	1.10	5.17	***	1.05
	Pre-Delayed	4.76	***	0.97	4.33	***	0.88	4.59	***	0.94
Cosine	Pre-Post	4.41	***	0.90	6.97	***	1.42	5.35	***	1.09
	Pre-Delayed	5.47	***	1.12	4.66	***	0.95	4.34	***	0.89
Logarithmic	Pre-Post	1.50	*ns*		1.37	*ns*		2.46	*	0.50
	Pre-Delayed	1.52	*ns*		2.22	*	0.45	2.90	*	0.59
CF	Pre-Post	0.64	*ns*		1.44	*ns*		2.07	+	0.42
	Pre-Delayed	1.27	*ns*		3.92	***	0.80	2.84	*	0.58

Table J.1. Summary of accuracy gain analyses. For the p columns, *** indicates p < .001, ** indicates p < .01, * indicates p < .05, + indicates p < .10, *ns* indicates p > .10.

Trained Items (TI)

Raw accuracy

A 3 phase x 3 condition ANOVA on TI accuracy showed a main effect of phase, $F(2, 138) = 100.82$, $p < .001$, $\eta^2_p = .59$. All conditions led to strong and persisting learning gains on trained items (pre vs. immediate posttest, $t(71) = 14.47$, $p < .001$, $d = 2.14$, pre vs. delayed test, $t(71) = 9.47$, $p < .001$, $d = 1.52$). This was the case for all conditions (*Table J.1*). All conditions improved from pretest to immediate posttest and from pretest to delayed test, p's < .001, d's = .99 to 1.97. There was also a reliable drop in accuracy from post to delayed test, $t(71) = 4.67$, $p < .001$, $d = .65$.

There was a main effect of condition, $F(2,69) = 3.52$, $p < .05$, $\eta^2_p = .09$. Both of the comparison conditions led to greater overall TI accuracy than the *AL* condition (43% from *AL/AC* vs. 35% from *AL*, $t(46) = 2.61$, $p < .05$, $d = .75$, and 41% from *AL/NC* vs. *AL*, $t(46) = 2.02$, $p = .05$, $d = .58$. There was no phase x condition interaction, $F(4,138) = .73$, $p > .10$.

Accuracy gain

The average accuracy gain was numerically higher for the *AL/AC* condition (35%) than 25% from the *AL* condition and 31% from the *AL/NC* condition, though the main effect was just marginally significant, $F(2,68) = 2.34$, $p = .10$, $\eta^2_p = .06$, and the pairwise t-tests showed no reliable differences among the three conditions, p's > .10. The greater the pretest TI, the smaller the learning gain at immediate posttest, $r(72) = -.32$, $p < .01$, and at delayed test, $r(72) = -.49$, $p < .001$. The difference between the immediate posttest gain and delayed test gain was not statistically significant, $F(1,68) = 2.10$, $p > .15$, confirming the notable retention over the one

week delay. There was no phase x condition interaction, nor phase x pretest accuracy interaction, *p*'s > .10.

The patterns were slightly different between Exponential items and Sine items. In both cases, *AL/AC* triumphed. *Figures J.1a and b* show the average accuracy on Sine and Exponential TIs.

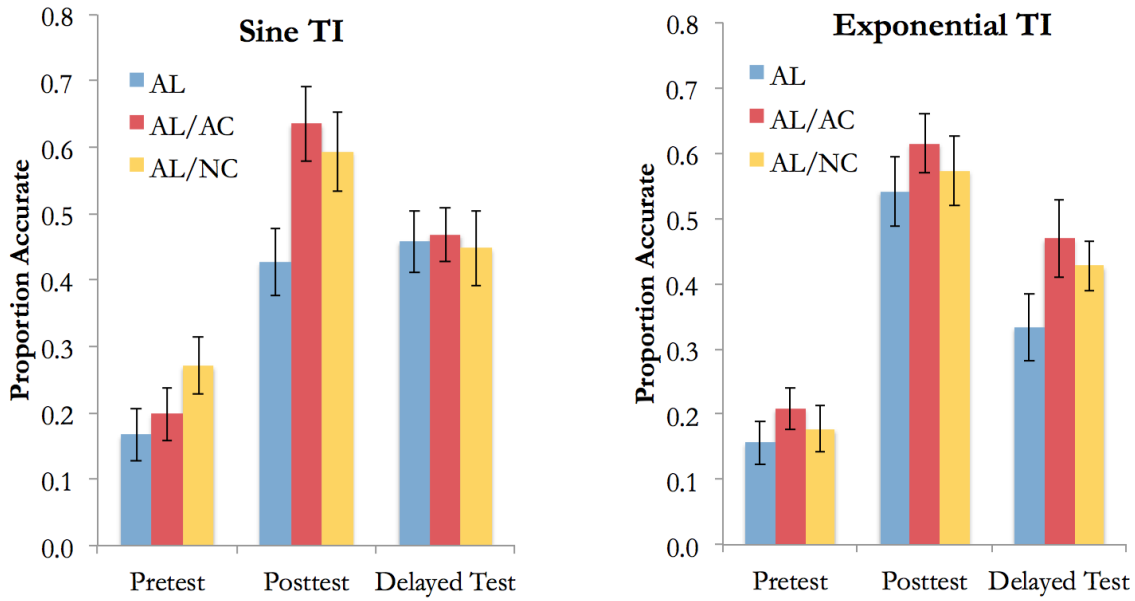


Figure J.1. Mean accuracy of (a) Sine trained items and (b) Exponential trained items

Sine TI

Raw accuracy. There was again a main effect of phase, $F(2,138) = 50.06$, $p < .001$, $\eta^2_p = .42$. The drop from immediate posttest to delayed test was modest, $t(71) = 2.57$, $p < .05$, $d = .36$, and there was large gains from pretest to immediate posttest, $t(71) = 9.34$, $p < .001$, $d = 1.38$, and to delayed test, $t(71) = 7.23$, $p < .001$, $d = 1.13$. There was a marginal phase x condition interaction, $F(2,138) = 2.13$, $p = .08$, $\eta^2_p = .06$, but the main effect of condition did not reach statistical significance, $F(2,69) = 2.15$, $p = .13$, $\eta^2_p = .06$. There were no condition differences at pretest, *p*'s >.10, but at immediate posttest, both of the comparison conditions did numerically

better than the *AL* condition (64% from *AL/AC* and 43% from *AL*, $t(46) = 2.74$, $p < .10$ (marginal), $d = .79$, and 59% from *AL/NC* vs. *AL*, $t(46) = 2.12$, $p < .05$, $d = .61$. There was no difference between the two comparison conditions, $p > .10$.

Accuracy gain. Similarly, the 2 x 3 ANCOVA with pretest Sine TI accuracy as the covariate confirmed an marginal phase x condition interaction, $F(2,68) = 2.83$, $p = .07$, $\eta^2_p = .08$, and no main effect of condition, $p > .10$. There was also no main effect of phase, or phase x pretest Sine TI interaction, p's > .10. The *AL/AC* condition had higher pre-post gain on Sine TI than the *AL* condition, $t(46) = 2.24$, $p < .05$, $d = .65$, but there were no other condition differences on pre-post nor pre-delayed gain, p's >.10. Again, the pretest predicted immediate posttest gain, $r(72) = -.44$, $p < .001$, and delayed test gain, $r(72) = -.60$, $p < .001$, as shown by a main effect of pretest, $F(1,68) = 40.36$, $p < .001$, $\eta^2_p = .37$.

Exponential TI

Raw accuracy. There was a marginal main effect of condition, $F(2,69) = 2.84$, $p = .07$, $\eta^2_p = .08$. There was larger gain by *AL/AC* (43%) than *AL* (34%), $t(46) = 2.31$, $p < .05$, $d = .67$. There were no reliable differences between *AL/AC* and *AL/NC*, and between *AL/NC* and *AL*, p's > .10. Improvements were found across all conditions, $F(2,138) = 56.01$, $p < .001$, $\eta^2_p = .45$. There was a drop from immediate posttest to delayed test, $t(71) = 4.43$, $p < .001$, $d = .67$, but the learning gain from pretest to immediate posttest was reliable, $t(71) = 10.92$, $p < .001$, $d = 1.90$, and so was the gain from pretest to delayed test, $t(71) = 6.10$, $p < .001$, $d = 1.09$. There was no phase x condition interaction, $p > .10$.

Accuracy Gain. The 2 x 3 ANCOVA with pretest Exponential TI accuracy as the covariate confirmed the drop in learning gain from the immediate posttest to the delayed test, $F(1,68) = 6.21$, $p < .05$, $\eta^2_p = .08$, but the main effect of condition did not reach statistical

significance, $F(1,68) = 2.32$, $p = .11$, $\eta^2_p = .06$. There were no phase x pretest and no phase x condition interactions, p's > .10. As with other measures, the pretest Exponential TI performance negatively correlated with both the posttests, $r(72) = -.60$, $p < .001$, and delayed test Exponential TI, $r(72) = -.64$, $p < .001$.

Trained Functions/Novel Items (TF/NI)

Raw accuracy

Across conditions, participants robustly transferred their learning to new instances of trained transformations, $F(2, 138) = 70.65$, $p < .001$, $\eta^2_p = .54$ (pre vs. posttest, $t(71) = 12.05$, $p < .001$, $d = 1.74$, pre vs. delayed test, $t(71) = 7.61$, $p < .001$, $d = 1.22$, and post vs. delayed test, $t(71) = 4.95$, $p < .001$, $d = .63$).

This was the case for all conditions. All conditions improved from pretest to immediate posttest and from pretest to delayed test, p's < .01, d's = .73 to 1.91.

This transfer ability was particularly strong for those in the *AL/AC* group, $F(2,69) = 5.79$, $p < .01$, $\eta^2_p = .14$. They performed better overall than both the *AL*, $t(46) = 3.47$, $p < .01$, $d = 1.00$, and the *AL/NC* groups, $t(46) = 2.29$, $p < .05$, $d = .66$. There was no condition difference between the *AL/NC* and *AL*, $p > .10$, and no phase x condition interaction, $p > .10$.

Accuracy gain

The 2 x 3 ANCOVA with pretest TF/NI as the covariate confirmed the condition differences in overall learning gain, $F(2,68) = 5.87$, $p < .01$, $\eta^2_p = .15$. Pairwise t-tests were not statistically reliable, but the 8% difference between *AL/AC* and *AL/NC* was marginally significant with a medium effect size, $t(46) = 1.75$, $p = .09$, $d = .51$. The lower the pretest TF/NI, the higher the learning gain at immediate posttest, $r(72) = -.51$, $p < .001$, and at delayed test,

$r(72) = -.68$, $p < .001$. There was no main effect of phase, suggesting little learning loss between immediate and delayed test, $p > .10$, and no interactions, p's > .10.

These patterns were similar for both Exponential and Sine TF/NIs. *Figures J.2a and b* show the average accuracy on Sine and Exponential TF/NI items.

Exponential TF/NI

Raw Accuracy. There was a main effect of phase, $F(2,138) = 78.39$, $p < .001$, $\eta^2_p = .53$. Regardless of conditions, there were strong improvements from pretest (19%) to immediate posttest (59%), $t(71) = 12.39$, $p < .001$, $d = 1.86$, and to delayed test (43%), $t(71) = 7.27$, $p < .001$, $d = 1.11$. The drop between immediate posttest and delayed test was reliable, $t(71) = 5.24$, $p < .001$, $d = .63$. There was a main effect of condition, $F(2,69) = 4.23$, $p < .05$, $\eta^2_p = .11$. The *AL/AC* group did better than *AL* (48% vs. 36%, $t(46) = 2.62$, $p < .05$, $d = .76$), and also better than *AL/NC* (38%, $t(46) = 2.40$, $p < .05$, $d = .69$). There was no phase x condition interaction, $p > .10$.

Accuracy gain. The drop in accuracy between immediate posttest and delayed test gain was confirmed with a main effect of phase, $F(1,68) = 6.21$, $p < .05$, $\eta^2_p = .08$. Pretest Exponential TF/NI strongly predicted immediate posttest and delayed test, $F(1,68) = 81.00$, $p < .001$, $\eta^2_p = .54$. The smaller the pretest, the greater the gain at immediate posttest, $r(72) = -.46$, $p < .001$, and delayed test, $r(72) = -.46$, $p < .001$. The main effect of condition did not reach statistical significance, $F(2,68) = 2.32$, $p = .11$, $\eta^2_p = .06$, and there were no interactions, p's > .10.

Sine TF/NI

Raw Accuracy. Similarly, all groups had strong improvements across learning phase, $F(1,138) = 20.76$, $p < .001$, $\eta^2_p = .23$. Participants scored 25% at pretest and 49% at immediate posttest, $t(71) = 6.15$, $p < .001$, $d = .92$, and remembered much of what they have learned at

delayed test with 42%, $t(71) = 4.16$, $p < .001$, $d = .72$. They forgot a small, albeit reliable amount between immediate posttest and delayed test, $F(71) = 2.06$, $p < .05$, $d = .30$. There was a main effect of condition, $F(2,69) = 3.12$, $p = .05$, $\eta^2_p = .08$. *AL/AC* did better than *AL* (44% vs. 35%), $t(46) = 2.49$, $p < .05$, $d = .72$, and there were no other condition differences, p's > .10, and no phase x condition interaction, $p > .10$.

Accuracy gain. A 2 phase x 3 condition ANOVA on the accuracy gain confirmed a main effect of phase, $F(1,69) = 4.24$, $p = .04$, $\eta^2_p = .06$, but there was no main effect of condition nor phase x condition interaction, p's > .10.

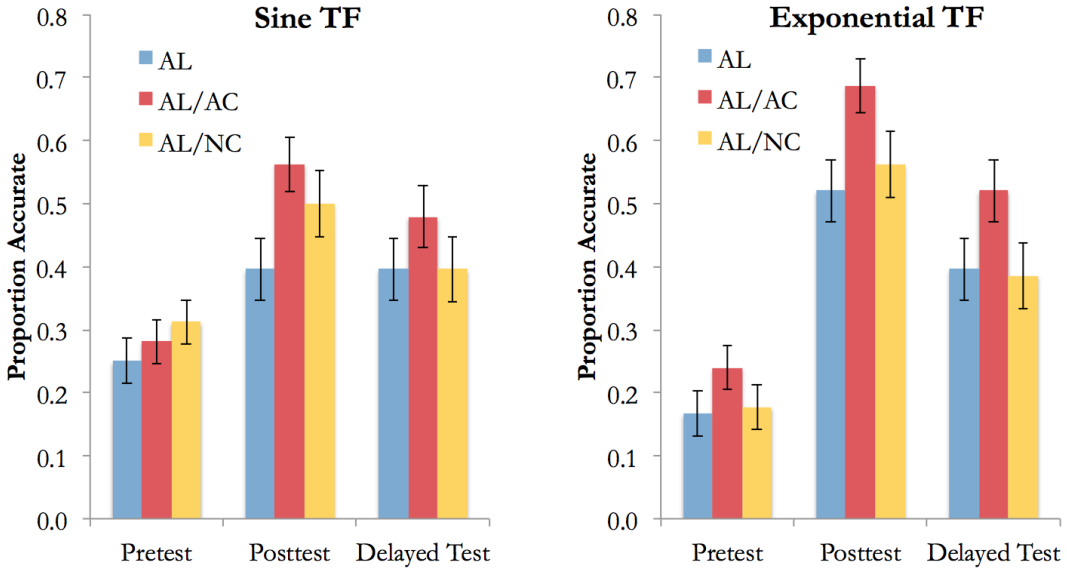


Figure J.2. Mean accuracy of (a) Sine and (b) Exponential trained functions, novel items

Untrained Functions (UF)

Raw accuracy

Figure J.3 shows the average accuracy on UF items. Training on Sine and Exponential functions also led to strong and long-lasting improvements on Cosine and Logarithmic functions,

but there were no differences among conditions on the amount of UF gained. The 3 phase x 3 condition ANOVA confirmed a main effect of phase, $F(2,139) = 41.60$, $p < .001$, $\eta^2_p = .38$. There was an 18% gain from pretest to immediate posttest, $t(71) = 8.31$, $p < .001$, $d = 1.13$, and 19% gain from pretest to delayed test, $t(71) = 7.88$, $p < .001$, $d = 1.18$. What was learned was retained a week later. The immediate posttest and delayed test UF accuracies did not differ from each other, $p > .10$.

This was the case for all conditions. All conditions improved from pretest to immediate posttest and from pretest to delayed test, p's $< .01$, $d = .78$ to 1.10.

The *AL/AC* group had higher average UF accuracy (32%) than both the *AL/NC* (26%) and the *AL* (27%) groups, but the main effect of condition was not statistically significant, $p > .10$. There was also no phase x condition interaction, $p > .10$.

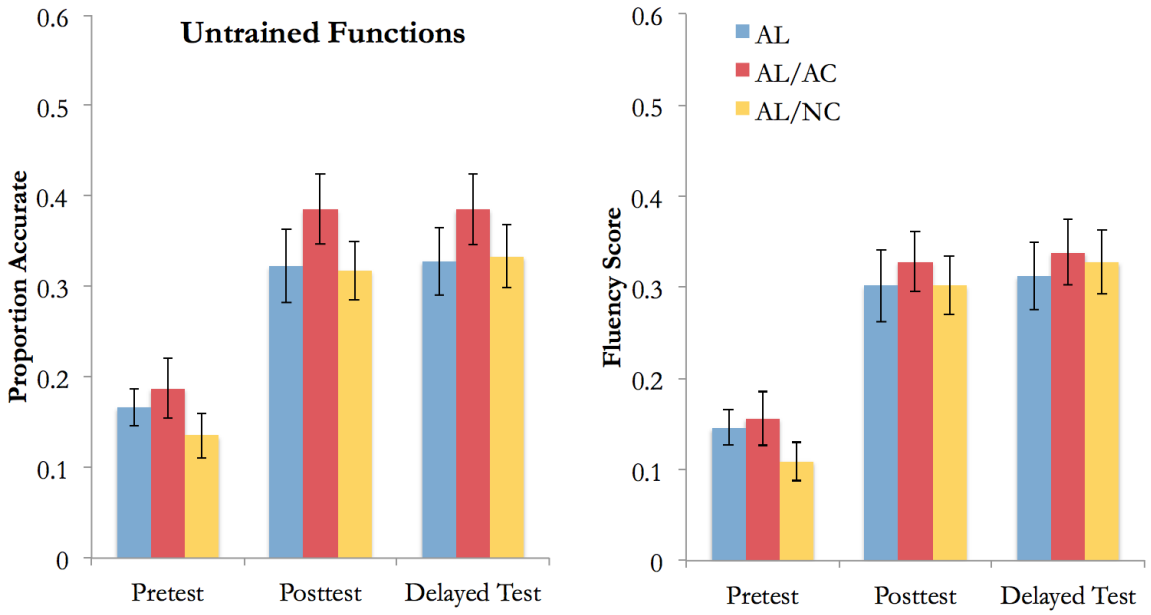


Figure J.3. Mean accuracy of untrained function items

Accuracy gain

A 2 phase x 3 condition ANCOVA confirmed that the lower the pretest UF accuracy, the higher the learning gain at immediate posttest, $r(72) = -.36$, $p < .01$, and delayed test, $r(72) = -.46$, $p < .001$. There were also no differences in accuracy gains at post and delayed test, no differences across condition in accuracy gains, and no phase x condition interaction, p's > .10.

Interestingly, there were condition differences on Cosine items but not on Logarithmic items. *Figure J.4* shows the average accuracy on Cosine and on Logarithm items.

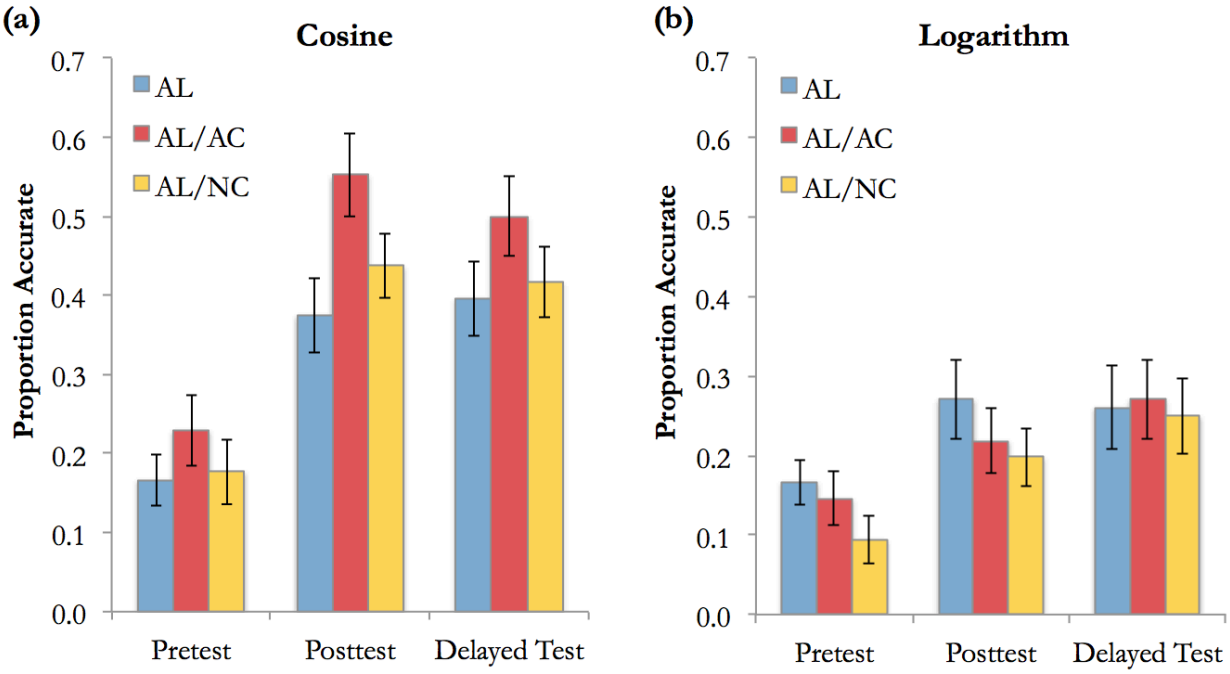


Figure J.4. Mean accuracy of (a) Cosine and (b) Logarithm untrained function items

Cosine

Raw accuracy. There was a marginally significant main effect of condition, $F(2,69) = 3.03$, $p = .06$, $\eta^2_p = .08$. The *AL/AC* condition did better than the *AL* condition, 43% vs. 31%, $t(46) = 2.21$, $p < .05$, $d = .64$, and marginally than the *AL/NC* condition, 43% vs. 34%, $t(46) = 1.74$, $p = .09$, $d = .50$, both with medium effect sizes. There was no difference between the *AL*

and the *AL/NC* condition, $p > .10$. As above, there was a main effect of phase, $F(2,138) = 51.06$, $p < .001$, $\eta^2_p = .43$, and no phase x condition interaction, $F(4,138) = .66$, $p > .10$. There was strong learning gain from pretest to immediate posttest, 19% to 45%, $t(71) = 9.31$, $p < .001$, $d = 1.21$, and to delayed test (44%), $t(71) = 8.37$, $p < .001$, $d = 1.15$, with no decay between immediate posttest and delayed test, $t(71) = .59$, $p > .10$. This learning gain was found with all conditions as seen in *Table J.1*, p's < .001, d = .89 to 1.42.

Accuracy gain. There was no difference in the pre-post than the pre-delayed gain, confirming long-lasting learning gain, $F(1,68) < 1$, $p > .10$. There was a marginally significant main effect of condition on the overall gain, $F(2,68) = 2.57$, $p = .08$, $\eta^2_p = .07$, but no reliable condition differences, p's > .10. There was a main effect of pretest on the learning gain, with pretest correlating highly with the immediate posttest gain, $r(72) = -.42$, $p < .001$, and delayed test gain, $r(72) = -.48$, $p < .001$.

Logarithm

Raw accuracy. There was a main effect of phase, $F(2,138) = 8.31$, $p < .001$, $\eta^2_p = .11$. Similar to Exponential items, across conditions, participants improved from pretest to immediate posttest, 14% to 23%, $t(71) = 3.28$, $p < .01$, $d = .51$, and from pretest to delayed test (26%), $t(71) = 3.90$, $p < .001$, $d = .62$, with no decay (but a small numerical gain) after a week, $t(71) = .92$, $p > .10$. There was no main effect of condition and no phase x condition interaction, p's > .10. Interestingly, only the *AL/AC* and *AL/NC* conditions showed improvements on Logarithmic items, but only at delayed test (*Table J.1*), p's < .05, d = .45 and .59, respectively.

Accuracy gain. Similarly there was a strong correlation between pretest and immediate posttest, $r(71) = -.52$, $p < .001$, and delayed test, $r(72) = -.46$, $p < .001$. There were no main effect of condition nor interactions, p's > .10.

Combination Functions (CF)

Raw accuracy

Figure 6.9 (in main text) shows the average accuracy on CF. Remarkably, all three trainings also led to improvements on CF, $F(2,138) = 10.26$, $p < .001$, $\eta^2_p = .13$. The gain from pretest to immediate posttest was modest (6%) and not reliable, $t(71) = 1.63$, $p = .11$, $d = .27$, but the gain from pretest to delayed test was a good 17%, $t(71) = 4.79$, $p < .001$, $d = .70$. Interestingly, this gain from immediate posttest to delayed test was reliable, $t(71)= 2.77$, $p < .01$, $d = .39$. This was the same pattern seen in Experiment 4.

Interestingly, similar to the *mixed* condition from Experiment 4, training with comparison trials led to improvements on CF. The *AL/AC* did not produce a reliable learning gain from pretest to immediate posttest, $p > .10$, but the learning gain from pretest to delayed test was one of large effect size, $t(23) = 3.92$, $p < .001$, $d = .80$. The *AL/NC* condition all conditions improved marginally on CF from both pretest to immediate posttest, $t(23) = 2.07$, $p = .05$, $d = .42$, and from pretest to delayed test, $t(23) = 2.84$, $p = .01$, $d = .58$, with small-medium effect sizes. The *AL* condition did not show any improvements form pretest to immediate posttest, nor from pretest to delayed test, p's $> .10$.

The main effect of condition was not statistically significant, $F(2,69) = 2.20$, $p = .12$, $\eta^2_p = .06$, and there was no phase x condition interaction, $F(4,138) = 1.58$, $p > .10$, $\eta^2_p = .04$.

Accuracy gain

The 2 phase x 3 condition ANCOVA with pretest CF accuracy as a covariate confirmed a marginal main effect of condition, $F(2,68) = 2.34$, $p = .10$, $\eta^2_p = .06$. Both of the trainings with comparisons led to greater accuracy gain on CF than the *AL* training, with marginal significance

but medium effect sizes (*AL/AC* vs. *AL*, $t(46) = 1.87$, $p = .07$, $d = .54$, *AL/NC* vs. *AL*, $t(46) = 1.73$, $p = .09$, $d = .50$). Across conditions, the immediate posttest gain did not differ from the delayed test gain, $p > .10$, and there was no phase x condition interaction, $p > .10$. There was no difference in pre-post and pre-delayed learning gain, $p > .10$. The accuracy at pretest could reliably predict learning gain, $F(1,68) = 67.99$, $p < .001$, $\eta^2_p = .50$, and there was a marginally significant phase x pretest accuracy interaction, $F(1,68) = 3.66$, $p = .06$, $\eta^2_p = .05$. The lower the accuracy at pretest, the higher the learning gain at immediate posttest, $r(72) = -.66$, $p < .001$, and delayed test, $r(72) = -.58$, $p < .001$.

Fluency Accuracy

Fluency scores show very similar patterns as accuracy results. *Figures 6.2b – 6.9b* show the fluent accuracy means.

Trained Items (TI)

Figure 6.6b shows the average fluent accuracy on TIs. There was a main effect of phase, $F(2,138) = 84.78$, $p < .001$, $\eta^2_p = .55$, showing large learning gain from pretest to immediate posttest (17% vs. 50%), $t(71) = 13.76$, $p < .001$, $d = 2.14$, and to delayed test (17% vs. 38%), $t(71) = 8.67$, $p < .001$, $d = 1.40$. The difference between post and delayed test was also reliable, $t(71) = 4.46$, $p < .001$, $d = .71$.

There was a main effect of condition, $F(2,69) = 2.49$, $p = .09$, $\eta^2_p = .07$. The *AL/AC* condition ($M = .37$, $SD = .09$) had overall marginally higher fluent accuracy than the *AL* condition ($M = .31$, $SD = .11$), $t(46) = 1.91$, $p = .06$, $d = .55$. The *AL/NC* condition ($M = .37$, $SD = .10$) also did marginally better than the *AL* condition, $t(46) = 1.87$, $p = .07$, $d = .54$. There was

no difference between the two comparison conditions, $p > .10$, and no phase x condition interaction, $p > .10$.

Fluent Accuracy Gain

There was a main effect of phase, $F(1,68) = 4.67$, $p = .03$, $\eta^2_{p=} .06$, reflecting a drop from immediate posttest to delayed test. There was no main effect of condition, $F(2,68) = 1.56$, $p > .10$, $\eta^2_p = .04$, and no interactions, p's $> .10$.

These were no differences among conditions on fluent accuracy when considering Sine TI and Exponential TI separately, p's $> .10$.

Trained Functions, Novel Items (TF/NI)

Figure 6.5b shows the average fluent accuracy on TF/NI items. All conditions were able to apply what they have learned fluently to TF/NI items. There was a main effect of phase, $F(2,138) = 89.97$, $p < .001$, $\eta^2_p = .57$. All groups showed strong improvements from pretest to immediate posttest (17% vs. 49%), $t(71) = 12.70$, $p < .001$, $d = 1.86$, and to delayed test (17% vs. 40%), $t(71) = 9.00$, $p < .001$, $d = 1.45$. The change from post and delayed tests was also reliable, $t(71) = 3.98$, $p < .001$, $d = .55$. There was a main effect of condition, $F(2,69) = 3.43$, $p = .04$, $\eta^2_p = .09$. The *AL/AC* condition ($M = .40$, $SD = .09$) had overall higher fluent accuracy than the *AL* condition ($M = .32$, $SD = .09$), $t(46) = 2.84$, $p = .007$, $d = .82$. There were no other condition differences, p's $> .10$.

There was a marginal phase x condition interaction, $F(4,138) = 2.01$, $p = .09$, $\eta^2_p = .06$. There were no condition differences at pretest, p's $> .10$, but at immediate posttest, both of the *AL/AC* and *AL/NC* conditions did better than the *AL* condition (*AL/AC* vs. *AL*, $t(46) = 3.45$, $p = .001$, $d = 1.00$; *AL/NC* vs. *AL*, $t(46) = 2.19$, $p = .03$, $d = .64$). However, there were no differences at delayed test, p's $> .10$.

Fluent Accuracy Gain

There was a main effect of condition, $F(2,68) = 4.45$, $p < .05$, $\eta^2_p = .12$. The *AL/AC* did marginally better overall than the *AL* condition, $t(46) = 1.80$, $p = .08$, $d = .52$. There were no other condition differences, p's > .10, and no main effect of phase and no interactions, p's > .05.

There was a main effect of pretest, $F(1,68) = 61.42$, $p < .001$, $\eta^2_p = .48$. Lower pretest predicts higher immediate posttest, $r(72) = -.49$, $p < .001$, and higher delayed test, $r(72) = -.67$, $p < .001$.

Interestingly, these condition differences were more obvious with Sine TF/NI than Exponential TF/NI items.

Sine TF/NI

<u>*Fluent Accuracy.*</u> There was a main effect of phase, $F(2,138) = 25.85$, $p < .001$ $\eta^2_p = .27$, and a marginal main effect of condition, $F(2,69) = 2.68$, $p = .08$, $\eta^2_p = .07$, and no phase x condition interaction, $p > .10$. All three groups improved on Sine TF/NI from pretest to immediate posttest, $t(71) = 6.84$, $p < .001$, $d = 1.00$, and from pretest to delayed test, $t(71) = 5.14$, $p < .001$, $d = .84$, with no forgetting between post and delayed test, $p > .10$. *AL/AC* did better than *AL* ($M = .40$, $SD = .13$ versus $M = .32$, $SD = .12$, respectively), $t(46) = 2.22$, $p = .03$, $d = .64$. There was no other condition differences, p's > .10.

<u>*Fluent Accuracy Gains*</u>. There was a main effect of condition, $F(2,68) = 4.62$, $p = .01$, $\eta^2_p = .12$, and a main effect of pretest Sine TF/NI score, $F(1,68) = 146.90$, $p < .001$, $\eta^2_p = .68$, and no other effects, p's > .10. However, none of the pairwise comparisons among conditions were statistically reliable, p's > .10.

Exponential TF/NI

Fluent Accuracy. There was just a main effect of phase, $F(2,138) = 71.13$, $p < .001$, $\eta^2_p = .51$, and no other effects, p's > .10. All three groups improved from pretest to immediate posttest, $t(71) = 11.90$, $p < .001$, $d = 1.86$, and pretest to delayed test, $t(71) = 7.30$, $p < .001$, $d = 1.16$, with some forgetting between post and delayed test, $t(71) = 4.33$, $p < .001$, $d = .55$.

Fluent Accuracy Gains. There was just a main effect of phase, $F(1,68) = 9.64$, $p = .003$, $\eta^2_p = .12$, and a main effect of pretest Exponential TF/NI score, $F(1,68) = 32.94$, $p < .001$, $\eta^2_p = .33$, and no other effects.

UF

Raw scores

Figure J.3b shows the average fluent accuracy on UF items. There was a main effect of phase, $F(2,138) = 44.83$, $p < .001$, $\eta^2_p = .39$. The learning gains had large effect sizes: pretest to immediate posttest (14% - 31%), $t(71) = 8.42$, $p < .001$, $d = 1.19$, and pretest to delayed test (14% - 33%), $t(71) = 8.16$, $p < .001$, $d = 1.28$. There was no drop in score between post and delayed test, $t(71) = .71$, $p = .48$, and no main effect of condition nor phase x condition interaction, p's > .10.

Score gain

There was no main effects nor interactions, $p > .10$. The same patterns were found with both Cosine and Logarithm functions.

Combination Functions (CF)

Raw scores

Figure 6.9b shows the average fluent accuracy on Combination items. There was a main effect of phase, $F(2,138) = 12.56$, $p < .001$, $\eta^2_p = .15$. Interestingly, although there was no significant gain between pretest and immediate posttest, $t(71) = .28$, $p = .78$, $d = .04$, there was a

gain between pretest and delayed test, $t(71) = 4.08$, $p < .001$, $d = .59$, and between immediate posttest and delayed test, $t(71) = .45$, $p < .001$, $d = .63$, both with medium effect sizes.

There was a main effect of condition, $F(2,69) = 3.26$, $p < .05$, $\eta^2_p = .09$. *AL/NC* had overall higher accuracy than *AL/AC* (39% vs. 27%), $t(46) = 2.31$, $p = .03$, $d = .67$, and *AL* higher than *AL/AC* (38% vs. 27%), $t(46) = 2.21$, $p = .03$, $d = .64$. This was likely because the *AL* and *AL/NC* groups started out with higher accuracy on these items than the *AL/AC* group (38% and 31% vs. 22%, respectively, $t(46) = 2.20$, $p = .03$, $d = .63$ and $t(46) = 1.49$, $p > .10$) There was no difference between *AL/NC* and *AL*, $p > .10$. There was no phase x condition interaction, $p > .10$.

Interestingly, the *AL* condition showed no improvement from pretest to immediate posttest, and pretest to delayed test, $p > .10$, but both *AL/AC* and *AL/NC* showed reliable gains from pretest to delayed test with medium effect sizes, p's $< .05$, $d = .55$ and .61.

Score gain

There was a marginally significant main effect of phase, $F(1,68) = 4.00$, $p = .05$, $\eta^2_p = .06$, and no other effects, p's $> .10$.

Progression of Learning

By Quartiles

Figure 6.10 show the average accuracy, RTc, and fluent accuracy by quartile and by block, only on AL trials. Across all four quartiles, the *AL/AC* condition produced higher accuracy than the *AL* condition, with medium to large effect sizes, $t(46) > 1.84$, $p = .02$ to .07, $d = .53$ to .81. Interestingly, however, this seemed to have resulted from a speed-accuracy trade-off. The *AL/AC* condition consistently look longer to reach the correct answers on all four quartiles than the *AL* condition with medium effect sizes, $t(46) > 1.84$, $p = .03$ to .05, $d = .53$ to

.61. The *AL/AC* condition also did better than the *AL* condition in terms of score, but only reliably or marginally reliably so in the 1st , 2nd, and 4th quartiles, $t(46) > 1.96$, $p = .03$ to .06, $d = .58$ to .71.

AL/AC did not differ from *AL/NC* in accuracy in the first 2 quartiles, but they did marginally better than *AL/NC* in the latter two, $t(46) = 1.84$, $p = .07$, $d = .53$, and $t(46) = 2.52$, $p = .02$, $d = .73$, for the 3rd and 4th quartiles, respectively. *AL/AC* also look longer (or marginally longer) on RTc than AL*/NC* on the 1st, 3rd, and 4th quartiles, $t(46) > 1.75$, $p = .01$ to .09, $d = .51$ to .79. There were no differences between these two conditions on scores in any quartiles, p's > .10.

Between the *AL/NC* and *AL* conditions, *AL/NC* did better (and marginally better) on accuracy and score than AL on the 1st and 2nd quartile, $t(46) > 1.80$, $p = .04$ to .08, $d = .52$ to .62, but that advantage disappeared in the 3rd and 4th quartiles, p's > .10. There were no differences on RTc, $p > .10$.

By Blocks

These condition differences seemed apparent even during the first few blocks into the training. *Figure 6.10* shows these data by block. Because only *AL* trials are plotted, the *AL/NC* and *AL/AC* groups had fewer than 12 trials per block. At block 1, the *AL/AC* and *AL/NC* did not differ from *AL* in accuracy, p's > .10, but the *AL/AC* already had marginally higher RTc than *AL* and significantly higher than *AL/NC* (10.67 seconds vs. 8.10 seconds and 7.99 seconds, $t(46) = 1.98$, $p = .05$, $d = .57$ and $t(46) = 2.29$, $p = .03$, $d = .66$, respectively). This pattern held when we considered RT generally (RT on both correct and incorrect trials). At block 1, *AL/AC* generally took marginally longer than *AL* ($M = 10.67$, $SD = 3.99$ vs. $M = 8.10$ seconds, $SD = 4.93$,

respectively), $t(46) = 1.98$, $p = .05$, $d = .57$ and significantly longer than *AL/NC* ($M = 7.99$, $SD = 2.96$), $t(46) = 2.29$, $p = .03$, $d = .76$.

We also analyzed the accuracy (not RTc because there were very few accurate items) during the first 4 AL trials in Block 1, and found no statistical differences in accuracy among the conditions, p's > .10. *AL/AC* had higher RT ($M = 11.54$, $SD = 6.10$) than *AL/NC* ($M = 8.64$, $SD = 3.91$), $t(46) = 1.96$, $p = .06$, $d = .57$, but not higher than *AL*, $p > .10$. Taken together, the lack of differences in accuracy and RTc earlier on in the training suggested that condition effects were due to the training. However, the *AL/AC* group generally spent longer on each trial, suggesting that some of the effects seen during the training may be due to individual differences in overall speed.

References

Anderson, A. L., Ross, B. H., & Chin-Parker, S. (2002). A further investigation of category learning by inference. *Memory & Cognition*, *30*(1), 119-128.

Andrews, J. K., Livingston, K. R., & Kurtz, K. (2005). Improving category learning through the use of context items: Compare or contrast? In B. G. Bara, L. Barsalou, & M. Bucciarelli (Eds.), *Proceedings of the 27th Annual Conference of the Cognitive Science Society* (pp. 121–126). Austin, TX: Cognitive Science Society.

Ankowski, A. A., Vlach, H. A., & Sandhofer, C. M. (2013). Comparison versus contrast: Task specifics affect category acquisition. *Infant and Child Development*, *22*(1), 1-23.

Ark, T. K., Brooks, L. R., & Eva, K. W. (2007). The benefits of flexibility: The pedagogical value of instructions to adopt multifaceted diagnostic reasoning strategies. *Medical Education*, *41*(3), 281-287.

Atkinson, R. K., Derry, S. J., Renkl, A., & Wortham, D. (2000). Learning from examples: Instructional principles from the worked examples research. *Review of Educational Research*, *70*(2), 181-214.

Biederman, I., & Shiffrar, M. (1987). Sexing day-old chicks: A case study and expert systems analysis of a difficult perceptual learning task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 640-645.

Bhalla, M. C., Mencl, F., Gist, M. A., Wilber, S., Zalewski, J. (2013). Prehospital electrocardiographic computer identification of ST-segment elevation myocardial infarction. *Prehospital Emergency Care*, *17*(2), 211–216.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. *From learning processes to cognitive processes: Essays in honor of William K. Estes*, *2*, 35-67.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123-144). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe and A. Shimamura (Eds.), *Metacognition: Knowing about Knowing* (pp. 185-205). Cambridge, MA: MIT Press.

Blackwell, L. S., Trzesniewski, K. H., & Dweck, C. S. (2007). Implicit theories of intelligence predict achievement across an adolescent transition: A longitudinal study and an intervention. *Child Development*, *78*(1), 246-263.

Bodemer, D., Faust, U. (2006). External and mental referencing of multiple representations. *Computers in Human Behavior, 22*, 27-42.

Boster, J. S., & Johnson, J. C. (1989). Form or function: A comparison of expert and novice judgments of similarities among fish. *American Anthropologist, 91*, 866-889.

Braithwaite, D. W., & Goldstone, R. L. (2014). Benefits of variation increase with preparation. *Proceedings of the Thirsty-Sixth Annual Conference of the Cognitive Science Society* (pp. 1940-1945). Quebec City, Canada: Cognitive Science Society.

Bryan, W. L., & Harter, N. (1899). Studies on the telegraphic language: The acquisition of a hierarchy of habits. *Psychological Review*, *6*(4), 345.

Carvalho, P. F., & Goldstone, R. L. (2014). The benefits of interleaved and blocked study: Different tasks benefit from different schedules of study. *Psychonomic Bulletin & Review*, 1-8.

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: Evidence from judgments of learning. *Psychonomic Bulletin & Review, 14*(1), 107-111.

Catrambone, R., & Holyoak, K. J. (1989). Overcoming contextual limitations on problem solving transfer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 5*(6), 1147-1156.

Chase, W. G. & Simon, H. A. (1973). Perception in Chess. *Cognitive Psychology, 4(*1), 55-81.

Chin-Parker, S., & Ross, B. H. (2002). The effect of category learning on sensitivity to within-category correlations. *Memory & Cognition*, *30*(3), 353-362.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioral Sciences, 2nd Edition.* Hillsdale: Lawrence Erlbaum.

Council of Chief State School Officers (CCSSO) and National Governors Association Center for Best Practices (NGA Center). (2010, June 2). Common Core State Standards for Mathematics.

De Jager, J., Wallis, L., & Maritz, D. (2010). ECG interpretation skills of South African emergency medicine residents. *International Journal of Emergency Eedicine*, *3*(4), 309-314.

Ducas, R. A., Wassef, A. W., Jassal, D. S., Weldon, E., Schmidt, C., Grierson, R., & Tam, J. W. (2012). To transmit or not to transmit: How good are emergency medical personnel in

detecting STEMI in patients with chest pain? *Canadian Journal of Cardiology*, *28*(4), 432-437.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363.

Estes, N. M. (2013). Computerized interpretation of ECGs supplement not a substitute. *Circulation: Arrhythmia and Electrophysiology*, *6*(1), 2-4.

Evered, A., Walker, D., Watt, A. A., & Perham, N. U. (2014). Discrimination training on paired cell images influences visual learning in cytopathology. *Cancer Cytopathology*, *122*, 200-210.

Fent, G., Gosai, J., & Purva, M. (2014). Teaching the interpretation of electrocardiograms: Which method is best? *Journal of Electrocardiology,* 190-193.

Gentner, D., & Gunn, V. (2001). Structural alignment facilitates the noticing of differences. *Memory and Cognition, 29*(4), 565-577.

Gentner, D., & Markman, A. B. (1994). Structural alignment in comparison: No difference without similarity. *Psychological Science, 5,* 152–158.

Gentner, D., Loewenstein, J., & Thompson, L. (2003). Learning and transfer: A general role for analogical encoding. *Journal of Educational Psychology, 95*, 393-405.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive Science, 7*, 155-170.

Gentner, D. (2005). The development of relational category knowledge. In L. Gershkoff- Stowe, & D. H. Rakison (Eds.) *Building object categories in developmental time.* Mahway, NJ: Lawrence Erlbaum Associates.

Gentner, D. (2010). Bootstrapping the mind: Analogical processes and symbol systems. *Cognitive Science*, *34*(5), 752-775.

Gibson, E. J., & Pick, A. D. (2000). *An ecological approach to perceptual learning and development.* New York: Oxford University Press.

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

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15*, 1–38.

Gick, M. L., & Paterson, K. (1992). Do contrasting examples facilitate schema acquisition and analogical transfer? *Canadian Journal of Psychology, 46*, 539- 550.

Givvin, K. B., Stigler, J. W., & Thompson, B. J. (2011). What community college developmental mathematics students understand about mathematics, Part II: The interviews. *The MathAMATYC Educator, 2*(3), 4–18.

Gillespie, N. D., Brett, C. T., Morrison, W. G., & Pringle, S. D. (1996). Interpretation of the emergency electrocardiogram by junior hospital doctors. *Journal of Accident & Emergency Medicine, 13*(6), 395-397.

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

Goldstone, R. L., Landy, D., & Brunel, L. C. (2011). Improving perception to make distant connections closer. *Frontiers in Psychology*, *2*, 385.

Goldstone, R. L., Landy, D., and Son, J. Y. (2008). A well grounded education: The role of perception in science and mathematics. In M. De Vega, A. Glenberg, and A. Graesser (Eds.), *Symbols, Embodiment, and Meaning* (pp. 327–355). Oxford Press.

Hammer, R., Hertz, T., Hochstein, S., & Weinshall, D. (2009). Category learning from equivalence constraints. *Cognitive Processing, 10*(3), 211–232.

Hampton, J. A., Estes, Z., & Simmons, C. L. (2005). Comparison and contrast in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*, 1459–1476.

Hayes, J. R. (1985). Three problems in teaching general skills. In S. F. Chipman, J. W. Segal, & R. Glaser (Eds.), *Thinking and learning skills: Research and open questions* (Vol. 2, pp. 391-4050). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hendrickson, A. T., Kachergis, G., Gureckis, T. M., & Goldstone, R. L. (2010). Is categorical perception really verbally mediated perception? In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 1216-1221). Austin, TX: Cognitive Science Society.

Higgins, E. J., & Ross, B. H. (2011). Comparisons in category learning: How best to compare for what. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the 33rd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Hoffman, A. B., & Murphy, G. L. (2006). Category dimensionality and feature knowledge: When more features are learned as easily as fewer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32,* 301–315.

Holyoak, K. J. (2005). *Analogy*. In K. J. Holyoak, & R. G. Morrison (Eds.), The Cambridge handbook of thinking and reasoning (pp. 117-142). New York: Cambridge University Press.

Hughson, A. L., & Boakes, R. A. (2009). Passive perceptual learning in relation to wine: Short-term recognition and verbal description. *The Quarterly Journal of Experimental Psychology, 62*(1), 1-8.

Hsu, A. S., & Griffiths, T. E. (2010). Effects of generative and discriminative learning on use of category variability. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.

Jablonover, R. S., Lundberg, E., Zhang, Y., & Stagnaro-Green, A. (2014). Competency in electrocardiogram interpretation among graduating medical students. *Teaching and Learning in Medicine*, *26*(3), 279-284.

Jacoby, L. L., Toth, J. P., Lindsay, D. S., & Debner, J. A. (1992). Lectures for a layperson: Methods for revealing unconscious influences. In R. Bornstein & T. Pittman (Eds.), *Perception without awareness* (pp. 81-120). New York: Guilford Press.

James W. (1890*). The principles of psychology*. New York: Holt.

Joy Cumming, J., & Elkins, J. (1999). Lack of automaticity in the basic addition facts as a characteristic of arithmetic learning problems and instructional needs. *Mathematical Cognition*, *5*(2), 149-180.

Kalish, C. W., & Lawson, C. A. (2007). Negative evidence in inductive generalization. *Thinking and Reasoning, 13*, 394–425.

Kalyuga, S., & Sweller, J. (2004). Measuring knowledge to optimize cognitive load factors during instruction. *Journal of Educational Psychology*, *96*(3), 558.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509–539.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist*, *38*(1), 23-31.

Karpicke, J. D., & Roediger, H. L. (2007). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*(2), 151-162.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*(6018), 772-775.

Kellman, P. J. (2002). Perceptual learning. (3rd Ed.). In H. Pashler, & C. R. Gallistel (Eds.), *Stevens' Handbook of Experimental Psychology* (Vol. 3, pp. 259–299). New York: John Wiley & Sons.

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

Kellman, P. J., & Massey, C. M. (2013). Perceptual learning, cognition, and expertise. *Psychology of Learning and Motivation, 58,* 117 – 165.

Kellman, P. J., Massey, C. M., Roth, Z., Burke, T., Zucker, J., Saw, A., et al. (2008). Perceptual learning and the technology of expertise: Studies in fraction learning and algebra. *Pragmatics and Cognition, 16*(2), 356–405.

Kellman, P. J., Massey, C., & Son, J. Y. (2009). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. *Topics in Cognitive Science*, 1-21.

Kok, E. M., de Bruin, A. B. H., Robben, S. G. F. and van Merriënboer, J. J. G. (2012). Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology. *Applied Cognitive Psychology, 66*, 854–862.

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

Kotovsky, L., & Gentner, D. (1996). Comparison and categorization in the development of relational similarity. *Child Development*, *67*, 2797–2822. doi: 10.2307/1131753

Krasne, S., Hillman, J. D., Kellman, P. J., & Drake, T. A. (2013). Applying perceptual and adaptive learning techniques for teaching introductory histopathology. *Journal of Pathology Informatics*, *4*.

Krasne, S., Stevens, C. D., Kellman, P. J., & Niemann, J. T. (under review). Mastering ECG interpretation skills through a perceptual and adaptive learning module.

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: Gaze-tracking study. *Radiology*, *242*, 396–402.

Kurtz, K. J., & Boukrina, O. (2004). Learning relational categories by comparison of paired examples. In K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (pp. 756–761). Mahwah, NJ: Lawrence Erlbaum Associates.

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, *4*, 863.

Landy, D., & Goldstone, R. L. (2007). How abstract is symbolic thought? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(4), 720.

Lee, H. S., Betts, S., & Anderson, J. R. (2015). Not taking the easy road: When similarity hurts learning. *Memory & Cognition*, 1-14.

Levering, K. R., & Kurtz, K. J. (2015). Observation versus classification in supervised category learning. *Memory & Cognition*, 43, 266-282.

Loewenstein, J., Thompson, L., & Gentner, D. (1999). Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review, 6,* 586–597.

Love, B. C. (2002). Comparing supervised and unsupervised category learning. *Psychonomic Bulletin & Review, 9*, 829–835.

Markman, A. B., & Gentner, D. (1997). The effects of alignability on memory. *Psychological Science, 8*(5), 363-367.

Markman, A.B., & Ross, B. H. (2003). Category use and category learning. *Psychological Bulletin*, *129*, 592- 613.

Massey, C. M., Kellman, P. J., Roth, Z., & Burke, T. (2011). Perceptual learning and adaptive learning technology: Developing new approaches to mathematics learning in the class-room. In N. L. Stein (Ed.), *Developmental and learning sciences go to school: Implications for education*. New York: Taylor & Francis.

McDaniel, M.A., & Butler, A.C. (2012). A contextual framework for understanding when difficulties are desirable. In A.S. Benjamin (Ed.), *Successful remembering and successful forgetting: a Festschrift in honor of Robert A. Bjork*. London, UK: Psychology Press.

Mele, P. (2008). Improving electrocardiogram interpretation in the clinical setting. *Journal of Electrocardiology, 41*, 438–9.

Mettler, E., & Kellman, P. J. (2014). Adaptive response-time-based category sequencing in perceptual learning. *Vision Research*, *99*, 111-123.

Mettler, E., Massey, C. M., & 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.* Boston, MA: Cognitive Science Society.

Montgomery, H., Hunter, S., Morris, S., Naunton-Morgan, R., Marshall, R. M. (1994). Interpretation of electrocardiograms by doctors. *British Medical Journal, 309,* 1551-1552.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519-533.

Morrison, W. G., & Swann, I. J. (1990) Electrocardiograph interpretation by junior doctors. *Archives of Emergency Medicine, 7,* 108-10.

Miele, D. B., & Molden, D. C. (2010). Naive theories of intelligence and the role of processing fluency in perceived comprehension. *Journal of Experimental Psychology: General*, *139*(3), 535.

Nakagawa, S. (2004). A farewell to Bonferroni: The problems of low statistical power and publication bias. *Behavioral Ecology, 15*, 1044-1045.

Namy, L. L., & Gentner, D. (2002). Making a silk purse out of a sow's ears: young children's use of comparison in category learning. *Journal of Experimental Psychology - General, 131*, 5–15.

NCTM. (2000). *Principles and standards for school mathematics*. Reston, VA: NCTM.

Nunes, T. (1999). Mathematics learning as the socialization of the mind. *Mind, Culture, and Activity, 6,* 33–52.

Paas, G. W. C., & van Merrienboer, J. J. G. (1994).Variability of worked examples and transfer of geometrical problem solving skills: A cognitive-load approach. *Journal of Educational Psychology, 86*(1), 122–133.

Perneger, T.V. (1998). What's wrong with Bonferroni adjustments. *BMJ: British Medical Journal*, *316*(7139), 1236.

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.

Recker, M. M., & Pirolli, P. (1995). Modeling individual differences in students' learning strategies. *The Journal of the Learning Sciences*, *4*(1), 1-38.

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players. *Psychological Science, 12*, 48–55.

Renkl, A. & Atkinson, R. K. (2003) Structuring the transition from example study to problem solving in cognitive skills acquisition: A cognitive load perspective. *Educational Psychologist,* 38, 15–22.

Renkl, A., Atkinson, R. K., Maier, U. H., & Staley, R. (2002). From example study to problem solving: Smooth transitions help learning. *The Journal of Experimental Education*, *70*(4), 293-315.

Renkl, A., Atkinson, R. K., & Grobe, C. S. (2004). How fading worked solution steps works – A cognitive load perspective. *Instructional Science, 32*, 59-82.

Rittle-Johnson, B., & Star, J. R. (2009). Compared with what? The effects of different comparisons on conceptual knowledge and procedural flexibility for equation solving. *Journal of Educational Psychology, 101*, 529-544.

Rittle-Johnson, B., & Star, J. R. (2007). Does comparing solution methods facilitate conceptual and procedural knowledge? An experimental study on learning to solve problems. *Journal of Educational Psychology, 99*, 561–574.

Rittle-Johnson, B., Star, J. R., & Durkin, K. (2009). The importance of prior knowledge when comparing examples: Influences on conceptual and procedural knowledge of equation solving. *Journal of Educational Psychology, 101*(4), 836-852.

Roediger, H. L., & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*, 181–210.

Rosenthal, J. A. (1996). Qualitative descriptors of strength of association and effect size. *Journal of Social Service Research, 21*(4), 37-59.

Russell, A. A., & Kellman, P. J. (1998, March). Teaching automatic, 3-D recognition of chemical structures. Presented at the American Chemical Society Meeting, Dallas, Texas.

Salerno, S. M., Alguire, P. C., Waxman, H. S. (2003). Competency in interpretation of 12-lead electrocardiograms: summary and appraisal of published evidence. *Annals of Internal Medicine, 138,* 751–60.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84*(1), 1–66.

Schwartz, D. L., & Bransford, J. D. (1998). A time for telling. *Cognition and Instruction*, *16*(4), 475-5223.

Shah, A. P., & Rubin, S. A. (2007). Errors in the computerized electrocardiogram interpretation of cardiac rhythm. *Journal of Electrocardiology*, *40*(5), 385-390.

Shiffrin, R. M., & Lightfoot, N. (1997). Perceptual learning of alphanumeric-like characters. In R. L. Goldstone, P. G. Schyns, & D. L. Medin (Eds.) *The Psychology of Learning and Motivation, Volume 36*. San Diego: Academic Press. (pp. 45-82).

Siegel, M., & Misselt, A. (1984). Adaptive feedback and review paradigm for computer-based drills. *Journal of Educational Psychology*, *76*(2), 310–317.

Silva, A. B., & Kellman, P. J. (1999). Perceptual learning in mathematics: The algebra-geometry connection. In M. Hahn, & S. C. Stoness (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 683–688). Mahwah, NJ: Lawrence Erlbaum Associates.

Snow, R. E. (1992). Aptitude theory: Yesterday, today, and tomorrow. *Educational psychologist*, *27*(1), 5-32.

Soderstrom, N. C., & Bjork, R. A. (2013). Learning versus performance. *Oxford Bibliographies Online: Psychology.* New York: Oxford University Press.

Sowden P. T., Davies I. R. L., Roling, P. (2000). Perceptual learning of the detection of features in *x*-ray images: A functional role for improvements in adults' visual sensitivity. *Journal of Experimental Psychology: Human Perception and Performance, 26*, 379–390.

Stigler, J. W., Givvin, K. B., & Thompson, B. (2010). What community college developmental mathematics students understand about mathematics. *The MathAMATYC Educator, 10*, 4–16.

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.

Weitnauer, E., Carvalho, P. F., Goldstone, R. L., & Ritter, H. (2014). Similarity-based ordering of instances for efficient concept learning. *Proceedings of the 36th Annual Conference of the Cognitive Science Society* (p. 1760-1765). Mahwah, NJ: Lawrence Erlbaum Associates.

Wise, J. A., Kubose, T., Chang, N., Russell, A., & Kellman, P. (2000). Perceptual learning modules in mathematics and science instruction. In D. Lemke (Ed.), *Proceedings of the TechEd 2000 conference.* Amsterdam: IOS Press.

Wood, G., Batt, J., Appelboam, A., Harris, A., & Wilson, M. R. (2013). Exploring the impact of expertise, clinical history, and visual search on electrocardiogram interpretation. *Medical Decision Making*, *34*(1), 75-83.

Yamauchi, T., & Markman, A. B. (1998). Category learning by inference and classification. *Journal of Memory and Language, 39,* 124–148.

Yamauchi, T., & Markman, A. B. (2000). Inference using categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 776–795.