A continuous-time approach to experimental psychology

By

Aaron K. Cochrane

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Psychology)

at the

UNIVERSITY OF WISCONSIN – MADISON

2020

Date of final oral examination: 3 September 2020

The dissertation is approved by the following members of the Final Oral Committee:

C. Shawn Green, Associate Professor, Psychology
Vanessa Simmering, Assistant Professor, Psychology
Maryellen MacDonald, Professor, Psychology
Emily Ward, Assistant Professor, Psychology
Haley Vlach, Associate Professor, Educational Psychology

Acknowledgments

I'm infinitely grateful to the academic communities I have been fortunate enough to be a part of. To pretend my work is solely my own would be hubris. The beginnings of my academic career would not have been possible without the guidance and support of Erica Kleinknecht, Connor Principe, Dane Joseph, Alyson Burns-Glover, Aaron Greer, Cheleen Mahar, and Chris Wilkes. After arriving at UW-Madison my nascent career has been fostered by the faculty members who have been valuable mentors, collaborators and intellectual resources: Vanessa Simmering, Seth Pollak, Karen Schloss, Ed Hubbard, David Kaplan, and Joe Austerweil, among many others. My intellectual environment has likewise benefited from my work with and alongside Gillian Dale, Florian Kattner, Clint Jensen, Mohan Ji, Lauren Anthony, and many other postdocs as well as graduate students; I owe a great debt to you all. I also cannot express enough appreciation for the undergraduate students who have worked on my projects and conducted their own research under my supervision. I look forward to further collaborations. Last, and not least, I am incredibly grateful to C. Shawn Green, for your inhuman ability to help graduate students like me and to generally just get things done.

Table of Contents

Abstract

Humans thoughts and behaviors are embedded in a continuously changing world, and effectively engaging with that world involves constant adaptation and learning of new information and skills. Yet, although learning occurs continuously as we interact with the world, theories in psychology nearly always rely on experimental approaches and data analytic techniques that assume humans are unchanging over the course of psychological experiments. Rather than understanding psychological processes as being unchanging over the course of many measurements, I instead approach behavior as continuously changing in response to experience. This perspective allows for a better alignment between psychological theories and the inferences allowed by experimental methods. In this dissertation I demonstrate the utility of a time-continuous approach to experimental psychology using simulations of behavioral data as well as empirical studies of perception, attention, memory, intelligence, and social cognition. Last, I describe the statistical package I have written to facilitate a continuous-time approach to understanding thought and behavior.

Chapter 1. In consideration of changing psychological processes

Overview

The chapters in this dissertation present converging evidence for the motivation, appropriateness, implementation and evaluation of nonlinear models of continuous change in a range of psychological domains. Specifically I will establish how disregarding continuous changes over time, as is exceptionally common in most conventional approaches to understanding psychological processes (e.g., categorizing time into large chunks such as sessions or blocks), can result in erroneous or suboptimal inferences regarding the underlying processes. Instead, by considering psychological processes as continuously evolving over time, I am able to align computational implementations with their corresponding theoretical assumptions. Critically, I will argue that this basic idea is true across much of the psychological sciences and in support of this argument, I demonstrate the improved theoretical inferences facilitated by approaching change as a continuous function of time in domains as varied as perceptual, cognitive, and social psychology.

Chapters 1 and 2 review the role of process-level change in psychological theories, the associated implementations, and the identification of certain theoretical and empirical questions. Chapter 3 provides evidence from several studies of perceptual learning to support a continuous-time perspective on the changes occurring as learning progresses. Chapter 4 provides several demonstrations of particular insights into psychological processes that are facilitated by a learning-centered perspective. Chapter 5 examines studies of individual differences in learning and memory, with an emphasis on the specific theoretical questions that can be asked when approaching behavior as continuously changing due to learning. Chapter 6 addresses pragmatic concerns of behavioral researchers regarding a learning-centered perspective on behavior, with an emphasis on a statistical package I have written to facilitate the implementation of continuous-time models. Chapter 7 serves to summarize and conclude this dissertation.

Dynamics of behavior: Situatedness within timescales

Charles Darwin is often said to have written, "…*it is not the strongest species that survive, nor the most intelligent, but the ones most responsive to change*." Despite there being no evidence for the

quote prior to the 1960s, the idea appeals to one contemporary notion of Darwinism: Adaptability and learning are precursors to success in changing environments, with the human species being wildly successful in part due to our ability to adapt and learn and develop within complex environments and on timescales ranging from seconds to decades. Indeed, all thoughts and behaviors are situated within interconnected systems and interconnected timescales (K. M. Newell et al., 2001; Thelen & Smith, 1994). This idea is at the core of many disparate literatures within psychology, from those that examine changes over developmental timescales (i.e., years) to those that examine changes on the timescale of brain activity (i.e., milliseconds). Yet, the situatedness of every observed action within short-term and long-term influences and demands poses a fundamental challenge to the study of human thoughts and behaviors. Although core theories may allow for (or even predict) dynamics on the scale of milliseconds, minutes, days, or years, in many cases, various practicalities nonetheless frequently force studies to abstract away from, or average over, timescales of change.

Abstraction away from fully time-dependent change may be necessary in some cases (although I will argue that it is necessary less often than it occurs in psychology). However, there is no need for this abstraction to be implicit. Instead, theories making claims about humans (e.g., related to memory or perception) should carefully consider the extent to which cross-person variations, as opposed to within-person patterns, are the ideal subject matter (Bornstein et al., 2006; Molenaar, 2004). Making explicit the within-person time-dependence of theoretical positions may, for instance, elucidate the developmental trajectories that give rise to psychopathology (Karmiloff-Smith, 1998). A lifespan approach to psychological inquiry, rather than being pigeonholed as "developmental," has become increasingly important for theories in areas such as working memory (Alloway & Alloway, 2013; Cowan & Alloway, 2009; Simmering, 2016) as well as broader theories attempting to integrate findings from disparate psychological disciplines (Westermann et al., 2007).

On the opposite end of the timescale spectrum, many models address changes in processing on timescales ranging from milliseconds to seconds. The emphasis is well justified; everyday behaviors and neural processes occur on this timescale. Common neural measures such as MEG and EEG provide

intensive data on changes over the course of seconds or fractions thereof, with fMRI and fNRIS providing information about dynamics on a slightly longer timescale. Computational models of perception, attention and memory address how information is processed on similar timescales (Constantinidis & Klingberg, 2016; Reynolds & Heeger, 2009; Simmering & Spencer, 2008; Smith & Ratcliff, 2004).

Lifespan timescales and neural-processing timescales have each been subject to extensive research, with the brief descriptions above failing to do justice to either. However, there is an intermediate timescale – that of minutes to hours - that is almost certainly the most accessible to behavioral researchers. Large swaths of experimental work involve, for instance, "a session lasting approximately 30 minutes" or "three blocks of trials each lasting 15 minutes", etc. Critically, for the purposes of my dissertation, this is also arguably the timescale where the idea of time-dependent change is least considered. Instead, the "sessions" or "blocks" very frequently become the fundamental unit of analysis – with performance being averaged over the entirety of the session or block (with few exceptions, e.g., Harlow, 1949). As noted above, this type of averaging over measurements precludes any understanding of the changes that would happen within the window of measurements (e.g., it implicitly assumes a lack of learning or fatigue), an issue that I turn to in detail in the next section.

Stationarity vs nonstationarity in experimental psychology

Prior to considering instances of trial-timescale changes in psychological processes, it is worth more thoroughly discussing the ways that conventional methods often inherently instantiate incorrect assumptions regarding psychological processes of interest. Within the course of any given behavioral experiment, it is extremely common for trials of the same trial type (e.g., set size, stimulus strength, semantic category) to be repeated some number of times. Each of these trials are then treated as if they arose from the exact same generative process. As such, the data resulting from these trials is implicitly assumed to be independently and identically distributed (i.e., *iid*). Take for example, a hypothetical visual short-term memory experiment where participants complete 60 total trials and where, on each trial, they attempt to remember and recall the shapes of 6 items. Treating all 60 trials as if they were identical, by

calculating an overall percent accuracy, would serve to abstract away from the within-experiment timescale by marginalizing over the dimension of time.

Marginalizing over the dimension of time necessarily implies that a process is *stationary*. I am not using the term stationary in any strict or technical sense here. Instead I am referring to a stationary process as any process that gives rise to *iid* observations, while the nonstationarity I will consider involves processes that give rise to *iid* observations *conditional on some time-dependent trend*. That is, the process of theoretical interest may change systematically over time and thus the distribution(s) of the observed data will likewise change. Systematic time-dependent changes may occur in response to many mechanisms (e.g., fatigue, external pressures), but in this dissertation I will limit my examination of nonstationarity to monotonic improvements in performance (e.g., learning, "warm-up").

There is a conventionally applied implicit equivalence between dimensions of time (Molenaar, 2004). In the short-term memory example described above, an average percent correct assumes that a person's ability to complete the task remains stationary throughout all 60 trials. An important part of this assumption is that the first trial is *interchangeable* with the sixtieth trial. The implication of this assumption should be clear: the system is assumed to be ergodic, meaning that measuring the first trial once, then the second trial once, and so on through the sixtieth trial, is *equivalent to measuring the first trial sixty times*. It is likewise equivalent to *measuring the sixtieth trial sixty times.* Note that I referred to the psychological process as being potentially (non-)stationary, and the measurements as being potentially (non-)ergodic; I will simply use the term (non-)stationary to refer to both process and data going forward. Figure 1 provides an example of an artificial case wherein, despite all participants having identical mean accuracies on a memory task, time-dependent patterns show systematic differences between participants. In this case, one participant's performance is stable over time, while the other two participants exhibit two different patterns of nonstationarity. If a researcher had a theory that working memory varies by participants' age, for example, then the aggregated results would provide compelling evidence for a lack of difference between participants. Critically though, this inference would be unequivocally incorrect. The very act of aggregating trials would necessitate a commitment to the belief that the ability to

remember items was stationary and measurements were *iid*, thereby precluding the possibility of finding the types of differences that exist between participants in this example (see Chapter 2 for additional examples).

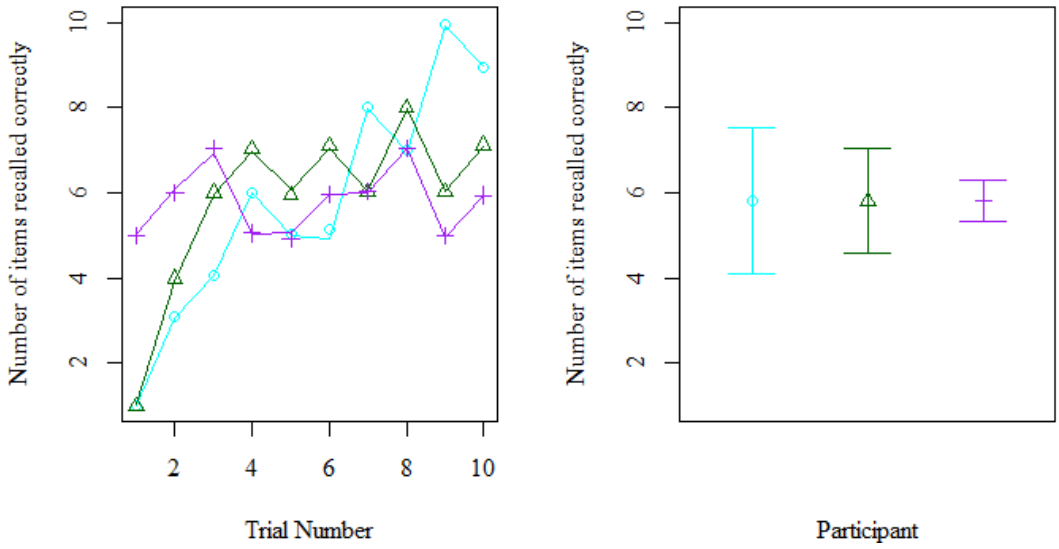


*Figure 1. Hypothetical example of the implications of stationarity assumptions. In the example working memory experiment described above, all participants have the same mean accuracy of 5.8 items (right panel; error bars = +/- 95% CI). Therefore, in a test of any theoretical frameworks where this score was used as the dependent variable, these three participants would be considered identical in terms of working memory ability. Yet, when examining performance through time (left panel), one participant appears stable over time, one participant appears to rapidly converge on stable scores, and one participant appears to be continuously changing. They are clearly not identical. Therefore, if one were to use the equivalent mean accuracy as evidence in support of a theoretical framework suggesting the three participants would be the same, it would be, at best, misleading. *Note that although variance is confounded with systematic change in this plot, the two features need not be confounded.*

To the extent that the generative process giving rise to the trials is truly unchanging, assuming stationarity is warranted and may be useful. However, this manuscript provides evidence that the assumption may not be justified even where it is conventionally made. As such, treating nonstationary processes as stationary may lead to noise or bias in inferences regarding the generative process. Even worse, important psychological processes that are involved within an experimental context may be completely disregarded when assuming stationarity. The potential for problems is most notable in the study of learning. Any time that a person is learning, their generative process is not stationary and performance at a given timepoint is therefore not interchangeable (i.e., *iid*) with their performance at other timepoints.

The roles of learning in theories of perception, attention, and cognition

Research in perception, attention, memory, and other cognitive processes is often conducted in manners that indicate a belief that the behaviors of interest are independent of learning-related change over time. This assumption of a lack of learning, and an associated assumption of process stationarity, are directly contradicted by influential theories in many domains. There is instead reason to treat learning as playing a central role in these domains. “Context effects,” “order effects” or, more generally, “top-down experience-based influences” are all ways of describing the results of learning (i.e., integration of previously-experienced information) in shaping behaviors.

Theories of low-level processing include compelling characterizations of perception as inference involving integration of current external information with established prior knowledge (Kersten et al., 2004; Ma, 2012). Indeed, to the extent that perception has top-down influences (e.g., goals, categories), these top-down signals must have come from prior experience that has accumulated over minutes, hours, or even years. Although the need for task-relevance and explicit feedback are debated within studies of the changes in perception with experience (i.e., perceptual learning; Seitz et al., 2006; Watanabe et al., 2001), the ubiquity of learning within basic perceptual processes is beyond question (Seitz & Dinse, 2007; Watanabe & Sasaki, 2015).

Theories of attention and memory likewise include learning effects. Nearly every theory of attention and memory, for instance, includes interactions between stimuli and prior experiences. For example, the effects of category distinctions in attention or "chunking" in short-term memory each necessitate learned information to be applied in the context of novel task demands or stimuli. Categories and "chunks" of information are flexibly adapted as well, providing evidence for interactions between longer-timescale and shorter-timescale learning processes (Casasanto & Lupyan, 2015; Cowan, 2001). In both attentional and decision-making processes there may be necessary links to perceptual learning processes providing enhanced extraction of information from the environment (Byers & Serences, 2012; Law & Gold, 2009, 2010).

Further, experimental psychology by design often necessitates learning. Participants tend to be placed into novel situations meant to mitigate the influence of previous experience (e.g., using ambiguous or randomly generated stimuli, requiring novel task demands, and pre-screening for certain types of language experiences or skills). Lessening the influence of prior experience putatively serves the purpose of decreasing unmeasured noise and increasing the extent to which in-experiment experience is the primary source of variation. If this is truly the case, within-experiment experience must necessarily include learning if participants are able to comply with task instructions and perform the task non-randomly. There are of course exceptions to this pattern, such as experiments in which participants are asked to complete impossible or trivial tasks while other dimensions of behavior or physiology are of primary interest. Nonetheless, much of experimental psychology requires participants to learn novel behaviors.

The fact that learning tends to occur within experimental psychology paradigms is not likely to be contested. For example, the ubiquity of "practice" trials is evidence for the implicit recognition of the presence of learning. More to the point, the removal of early trials belies a belief that implicitly-recognized learning is (1) irrelevant to the processes of true interest in any given study and (2) that any "nuisance" learning effects in the data can be eliminated by censoring the first N trials as "practice" (i.e., the assumption is that by removing those early trials, the time-dependent and nonstationary aspect of task

performance is removed, leaving the remaining trials *iid*). Rarely is a justification provided for learning being irrelevant (i.e., why the way that people learn a task doesn't speak to their underlying abilities or the underlying processes). Likewise, rarely is a justification provided for a particular quantity or type of learning (e.g., why choose 10 trials, or 20 trials, or particular types of trials to remove). The theoretical foundations of much of experimental psychology rely upon individuals' integration of current demands and information with prior learning, often constrained within the context of a given experimental setting. Given this relevance one might imagine that the deletion of early performance and ignoring learning should be the exception, rather than the rule. At a minimum it is surprising that deletion of data does not need justification, but is instead treated as convention.

The brief examples provided here clearly do not provide thorough or systematic evidence that learning is an important aspect of behavior to consider in all contexts. Such universality was not intended. However, overlapping timescales relevant to these theories uniformly include changes occurring over the course of minutes and hours. It should be clear that stationarity is an assumption that must be justified rather than simply implied. If context dependence is indeed a possibility in such a wide variety of domains, then the convention of implicitly assuming stationarity disregards this possibility.

Discretizing time as application of implicit theory.

Assumptions of stationarity

Rather than interrogating the possible time-dependent changes in psychological processes, it is not an understatement to say that the overwhelming majority of behavioral studies implicitly assume some stationarity of measurements. As addressed in several chapters here, treating measurements as if they arise from a stable psychological process could be a completely unsupported assumption. Although researchers may claim to be interested only in time-invariant stationary processes, this claim itself does not justify ignorance to the possibility of nonstationarity. Such a presupposition would be to commit, for example, to all patterns of performance in Figure 1 as being exactly equivalent from the perspective of a given theory. Such a commitment seems at best problematic, to the extent that within-person systematic differences can be understood. This point is reminiscent of arguments for multilevel models being more

powerful and less biased by accounting for within-participants random effects (Baayen et al., 2008; Moscatelli et al., 2012). In this vein, time-dependent changes can be thought of as a participant-level random effect that explains variation that would otherwise be unmodeled. Further, understanding time-evolving aspects of psychological processes may actually provide more leverage on theoretical questions than assuming stationarity (see Chapters 4 & 5).

Blocks of learning assume stationarity

Interestingly, assumptions of stationarity are not unique to the study of putatively "stable" processes like executive function, fluid intelligence, or working memory. Instead, these assumptions are implicitly implemented even in studies of learning itself. Despite a topic of study nominally concerned with the changes in knowledge or performance with each successive learning event (e.g., trial), dominant approaches to studying learning nonetheless nearly always involve categorizing time into discrete chunks (e.g., sessions or blocks). This is statistically problematic for the same reasons as any conversion of continuous data into categorial data (MacCallum et al., 2002). Categorizing some set of measurements as identical (e.g., as if they were all block 1 trials) blatantly assumes stationarity at the same time that the underlying theory of learning predicts change with successive learning events, thereby placing theory and inferential approach in direct opposition. Figure 2 demonstrates one such result, in which learning clearly varies between participants while aggregated analyses would conclude that all participants learned to the same extent. Aggregated analyses would therefore obscure systematic differences between the learning outcomes. This issue will be explored at length in Chapter 3, with several demonstrations of improved theoretical inferences and model fits when treating learning as occurring continuously through time.

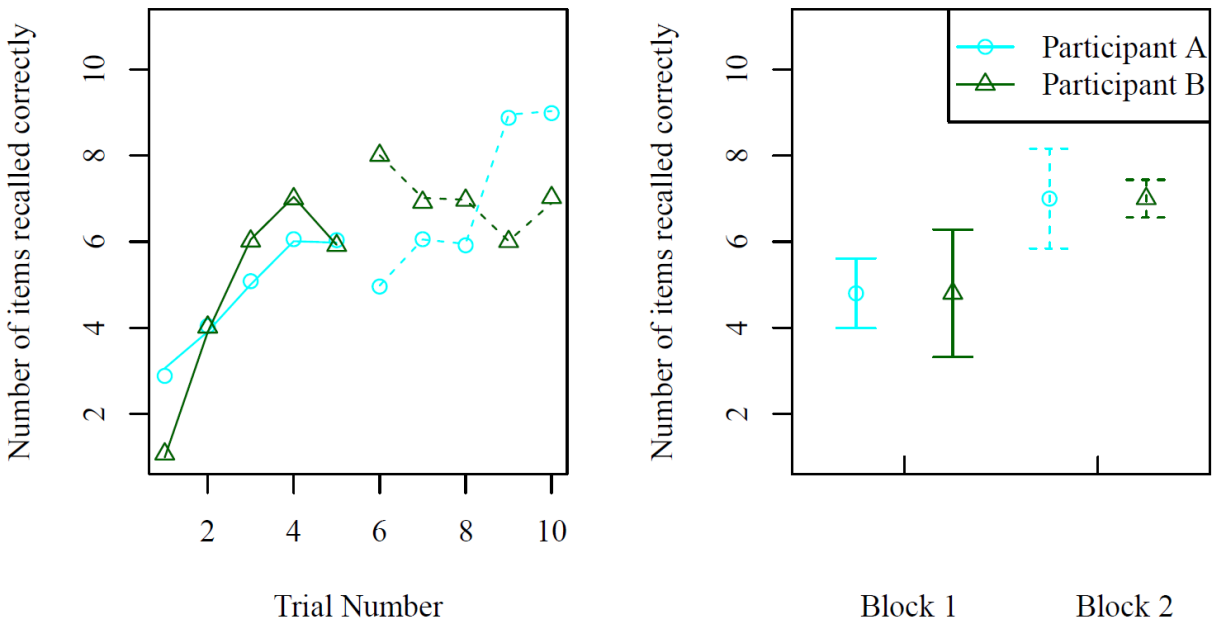


Figure 2. Hypothetical example of stationarity assumptions confounding learning effects. Ten experimental trials are split into two blocks of five trials each, as indicated by dashed or solid lines. Participants have equivalent averaged performance (right panel) in both "block 1" (4.8) and in "block 2" (7). These averages fail to convey that the two participants' performances are far from equivalent (left panel); Participant B appears to improve rapidly and then reach a stable level of performance, while Participant B improves throughout both blocks. These shapes of learning curves (i.e., rates of change) can convey valuable information about processes underlying learning (see Studies 3.2, 4.3c, 5.1, & 5.2).

Advanced models

Not all approaches understanding psychological processes involve assumptions of stationarity. Two common approaches, autocorrelation models and reinforcement learning, each can provide estimates of the relationships between adjacent timepoints. They thus naturally incorporate updates to knowledge and abilities on every learning event. It is even possible that certain modeling approaches in this manuscript may be replicable using reinforcement learning or autocorrelation models. I have limited the scope of my work in this dissertation, though, by approaching the investigation of learning using methods that are more directly comparable to conventional studies of perception and cognition. By extending

standard models (e.g., psychometric functions, signal detection theory) to incorporate nonlinear time-dependent dynamics I directly test the assumptions and implications of more conventional methods.

A way forward: avoiding untested assumptions by establishing empirical foundations

This dissertation is intended, in large part, to demonstrate a need for interrogating assumptions of stationarity in a variety of domains. In experiments ranging from visual perception to social cognition I question the stationarity of the processes concerned. In each case I demonstrate the substantial theoretical detriments of failing to consider the time-evolving nature of psychological processes. Many of the components of this dissertation could themselves be the backbones of entire independent theses. However, my explicit goal was to address a much broader point than that which could be made by an exclusive focus on any of the specific domains or experiment-types alone. To this end I actively avoid over-emphasizing any specific study within this dissertation because the importance of nonstationarity, and specifically of learning, cuts across many domains in behavioral sciences. To be too focused on a single example would obscure the larger importance. Even to focus on learning alone, eschewing considerations of other influential mechanisms of change such as fatigue, unfortunately limits the implications of this dissertation.

It is worth noting at the beginning of this dissertation that I am not claiming that all experiments must incorporate considerations of learning, fatigue, “warm-up” or other time-dependent trends. Nonstationarity may truly be a nuisance within certain processes of theoretical interest, but even in this case nonstationarity can be tested and controlled for. Whether or not change is of specific theoretical interest, an alternative change-centered approach to psychological processes necessitates a set of tools to identify stationary or nonstationary aspects of mental functioning, and to this end I have developed a statistical package that can seamlessly integrate common analyses in **R** with tests of assumptions of stationarity as well as implement time-related detrending of data (see Chapter 6). It is important to emphasize that, if a process is indeed stationary (i.e., has no time-evolving trends), these methods can demonstrate a lack of evidence for nonstationarity.

A fully change-centered approach to human psychology should be grounded in a wide base of evidence in both paradigms explicitly designed to study learning and those putatively independent of learning. Chapters 4 and 5 include various behavioral experiments as demonstrations of the improved theoretical inferences provided by decomposing data into time-evolving components. The active rejection of the assumption of stationarity involves utilizing observed change in the service of better understanding human behaviors. Tools to implement these inferences are also available in the **R** package that I have developed and describe in Chapter 6. Chapter 3 establishes, in studies of perceptual learning, the empirical and computational foundation upon which the later chapters rely.

Chapter 2 includes two sets of simulations to make the framework outlined in this chapter more concrete. I simulate data from a comparison of two groups completing a short-term memory span task. Limited inferences are possible given assumptions of stationary (i.e., averaging all performance) or an implicit assumption of learning (i.e., separating trials into “practice” and “real” data). In contrast, modeling trajectories of change provides specificity in understanding differences between the simulated groups.

Chapter 2. Simulations demonstrating effects of categorical vs. continuous time

As introduced in Chapter 1, theories of psychological functions tend to be founded on approaches that treat data as *iid* and therefore imply processes' stationarity and ergodicity (Molenaar, 2004). Here I will use two simulations of data to demonstrate several potential problems with the assumption of stationarity within a set of behavioral measurements. The simulations demonstrate two dimensions that may influence the inferences that can be/are made from a simple hypothetical study of visual short-term memory in which two groups' performance is compared. Simulated participants learn over the course of the task, which is fully in accordance with the conventional use of "practice trials" at the beginning of behavioral measurements.

The first dimension that may influence the inferences is, quite trivially, the actual state of the world. If true differences exist between the two groups being compared in a study, the analytic approach should lead to inferences supporting a difference between the groups. True differences arising from systematic variations in the generative processes giving rise to the data should align with the theory-grounded understandings of the very processes involved. If a participant completing a task is learning within the context of the task, as conveyed by the conventional use of "practice trials," one of the processes giving rise to the data would be learning. In this case, initial performance, rate of learning, and asymptotic performance may independently be indicative of processes that are informative to theory. The simulations demonstrate that variations in several aspects of learning trajectories can provide meaningful distinctions between various potential sources of variation in behavior.

The second dimension involves the various options that researchers have regarding the analysis of their data. To the extent that scientific inferences are constrained by the analytical approach to data, analyses allowing the disentanglement of plausible generative processes will allow process-level inferences that are impossible using analyses assuming *iid* data. When learning is a potentially influential process this issue becomes extremely clear. If learning is a suspected influence on performance the standard approach is to categorize time into discrete windows. Time may be categorized as "beginning"

and “end” if the topic of research is learning or else categorized as “practice” and “real data” if the topic is not learning itself. Both approaches necessitate an arbitrary choice to make time into a discrete, rather than continuous, dimension. For simplicity, the simulations in this chapter involve the common “practice” and “real data” categorization of the continuous dimension of time.

To the degree that both dimensions are related to some underlying theory, the generative process and the analysis of data should clearly be linked. As discussed after the simulations are reported, it is impossible to endorse any specific particular analytical procedure independently of the theoretical context. Rather, the simulations show reasons to consider interactions between possible psychological processes and feasible analytical approaches.

A third dimension influencing inferences in experimental psychology, that of methodological choices made by the researcher in data collection, is extremely important but is not considered directly here. In this chapter I briefly mention changes in aggregated results due to varying trial numbers. Parametric variations on methodological choices and their influence on the ability to make accurate inferences are discussed in more detail in Chapter 6.

It is important to note that the simulations reported here make several assumptions that provide for somewhat idealized scenarios for the identification of situations where aggregated analyses fail and continuous-time analyses do not. In particular, the knowledge of the exact learning function involved in generating the data provides the researcher with perfect knowledge regarding the parameterization of a model (see Study 3.3). The present simulations are therefore intended to be illustrative regarding cases in which aggregated analyses break down (i.e., the psychological process and the analytical process are at odds), while time-continuous analyses provide appropriate inferences regarding the data. In later chapters many empirical cases will be considered where, within some contexts considered to involve learning and other contexts that tend to assume stability, taking a time-continuous approach to inferences provides benefits.

Methods

Simulations of possible patterns of memory-task results

The data was generated as if from an experiment utilizing 64 trials of a memory span task of set size 6. On each trial the participant can answer between 0 and 6 items correctly, leading to single-trial accuracies of 0, .167, .333, and so on. Within-trial time dynamics (i.e., relating to the fact that the participant makes multiple responses in each "trial") were not considered here for the sake of simplicity; while nested timescales of change are a promising area of future inquiry, the current simulations fit each trial as providing a single data point.

I compared three types of analysis approaches – two typical approaches that do not (fully) take into account time as a factor and one fully time-dependent method. For the former (non-time-dependent) approaches, I considered both the reasonably standard approach of simply averaging all trials together into one task score as well as an approach where the first 16 trials were discarded as "practice" and the remaining 48 trials were averaged into a single task score (i.e., as the "real data"). A last alternative would be to model time-dependent change in accuracy. I varied simulation parameters to explore the implications of different generative processes (i.e., the psychological phenomena of interest) and analytical choices.

Fitting methods

For each demonstration, 40 participants were simulated with 64 trials total each, which is a realistic number in each dimension for actual cognitive psychology research. Two groups were simulated, with 20 participants each, and each group was defined as having a particular time-evolving accuracy on the task. The time-evolving accuracy was a saturating exponential function of time defined using a starting percent correct, an asymptotic percent correct, and a rate of change. (see Appendix 1 for code)

Individual participants' trials were sampled from the binomial distribution with a mean success rate equal to the group mean success rate, which changed over time as described above, and a number of possible successes equal to 6 (i.e., a "memory task set size of 6"). For example, if a group's accuracy for a given trial was .7, then on that trial each participant from the group had approximately a 6% chance of scoring 2 out of 6 (.33), 18.5% chance of getting 3 out of 6 (.50), 32.4% chance of getting 4 out of 6 (.67), 30.2% chance of getting 5 out of 6 (.83), and 11.7% chance of getting 6 out of 6 (1.0). The data was

modeled using a beta distribution, which is a continuous distribution allowing for values bounded by 0 and 1. As such, a lapse rate (i.e., edge correction) of .005 was applied to data before any analysis in order to ensure estimability by the beta distribution.

The R package **brms** was used to fit the time-evolving models (Bürkner, 2017). Time-evolving models were fit using an exponential function of change (Heathcote et al., 2000; A. Newell & Rosenbloom, 1981) parameterized in terms of the initial accuracy, final accuracy, and log time constant (i.e., number of trials to complete a certain percent of learning). Initial and final accuracies were estimated on inverse-logit scales such that parameter values could range from negative infinity to positive infinity while accuracy values would be bounded to 0 and 1. Given the parameterization, priors provided minimal information for the estimation of start (Gaussian with mean 0 and SD 2), rate (Gaussian with mean 8 and SD 4), or asymptote (Gaussian with mean 0 and SD 2). All other parameters used default priors. Each of the three parameters was estimated within the nonlinear model of change using a mixed-effects linear model with a fixed effect of group and a random intercept for each participant. Three sampling chains were run for 2500 iterations each for each simulation, with the final 1000 samples of each chain being retained for analysis. All R-hats were below 1.02, indicating model convergence.

The following two examples are particularly illustrative combinations of generative process [parameters] and methods [trial number and set size] such that standard analyses provide null or inconsistent results while time-dependent analyses reveal the differences between groups. However, it should be clear in the explanations that changes in any of these dimensions could change the results. Specifically, changing trial numbers and/or practice-trial exclusions can "create" non-null results, sometimes with opposite interpretations.

Results

Simulation 1: Aggregation provides inconsistent rejection of the null

The first simulated experiment demonstrates the challenges in using aggregated analyses to distinguish between variation in rates of change or asymptotic levels of performance (see Figure 3). In this case, the generative model starting accuracy does not differ between groups (both 0.4), learning rate

is different between groups (group 1: 5.75; group 2: 4.5), and asymptotic accuracy is different between groups (group 1: 0.89; group 2: 0.7).Yet, when making comparisons across groups using aggregated statistics, the true differences that exist between the groups in terms of learning and asymptote are obscured. This is true when data is aggregated across all trials (T(38) = 0.44) or when the first 16 trials are removed as practice and only the last 48 trials are aggregated across (i.e., standard analysis; T(38) = 0.78). In contrast, when using a learning model to fit performance, group-level differences show specific patterns (that, as expected, recover the true process that was used to create the data). The coefficient for the group difference in start is indistinguishable from zero (start = 0.04 [-0.23,0.3]), while the other two parameters show reliable group-level difference coefficients (rate = -1.28 [-2,-0.09]; asymptote = -1.83 [-3.6,-0.64]).

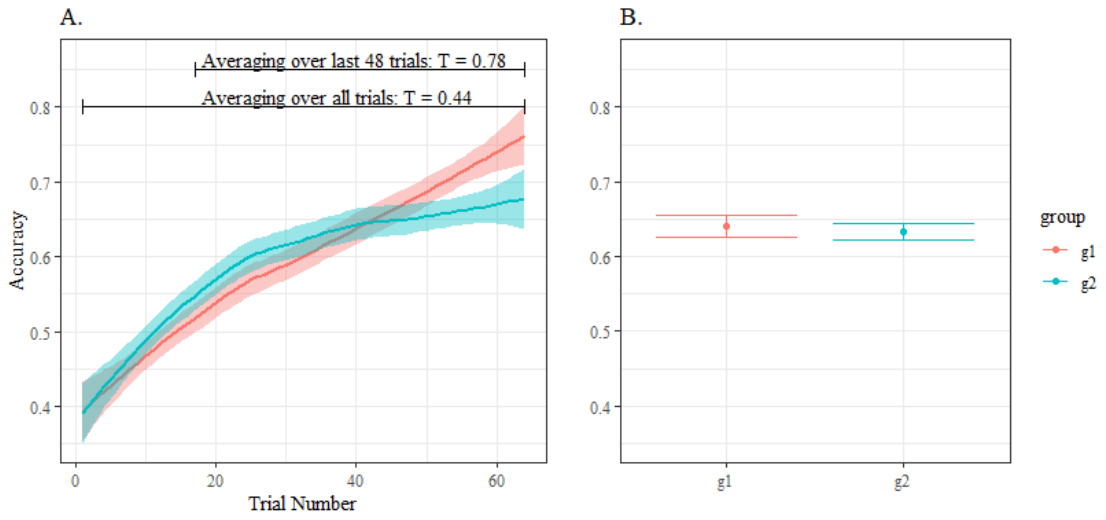


Figure 3. A simulated comparison between two groups in which the two groups have the same initial performance, but Group 1 has higher asymptotic accuracy than Group 2, while Group 2 shows faster learning than Group 1. As noted at the top of the figure, when taking averages across either the entire data set, or else when discarding 16 trials as practice, and then averaging over the remaining 48 trials, the true differences across groups are observed. (A) Lines are LOESS smoothed mean accuracies across

all participants within each group, with error bands indicating 95% CI of smoothed means. (B) Means and bootstrapped 95% CI of a standard analysis (i.e., T test comparing final 48 trials).

In this situation the two groups have identical true levels of accuracy at two points, once in the beginning and once at about trial 40. If more data were collected per participant and the first 40 trials were discarded as practice, then a between-group T test would indicate that Group 1 outperformed Group 2. In contrast, if fewer trials were collected, the opposite conclusion would be made. If only 40 trials were collected in total, then a T test would indicate that Group 2 outperformed Group 1. This latter conclusion would be made regardless of practice-trial exclusion. Overall, then, the choice of timescale would influence whether researchers would conclude that Group 1 performed better than Group 2, the groups were indistinguishable, or Group 1 performed worse than Group 2. This is obviously exceptionally problematic in terms of testing theory (i.e., if the choice of number of trials could create any of the three possible patterns of results in terms of comparing aggregated performance). Time-evolving analyses meanwhile are not influenced by the number of trials, provided of course that the function of time-related change appropriately matches the function of change in the generative process. The need for an appropriate match between functions justifies tests of the functional forms of change (see Chapter 3).

Simulation 2: Aggregation obscures effects of interest by emphasizing asymptote

In the second simulation, the two groups begin with different performance (Group 1: 0.4; Group 2: 0.5) and learn at different rates (Group 1: 4.4; Group 2: 5), yet they end with similar accuracies (both 0.8). As above, this simulated experiment demonstrates a data set in which, when aggregated, the true differences between fast learning and high starting accuracy are obscured (see Figure 4). In this model, there are no systematic differences between groups are evident when aggregated across all trials $T(38) = 0.43$) or when comparing scores aggregated [averaged] within individuals' last 48 trials (standard analysis $T(38) = -1.03$). In contrast, when using a learning model to characterize performance, group-level differences show specific patterns. The coefficient for the difference in asymptote does not meet the

criterion for reliability (1.19 [-0.1,3.95]), while the other two parameters show reliable group-level difference coefficients (rate = 1.21 [0.13,2.05]; start = 0.39 [0.15,0.63]).

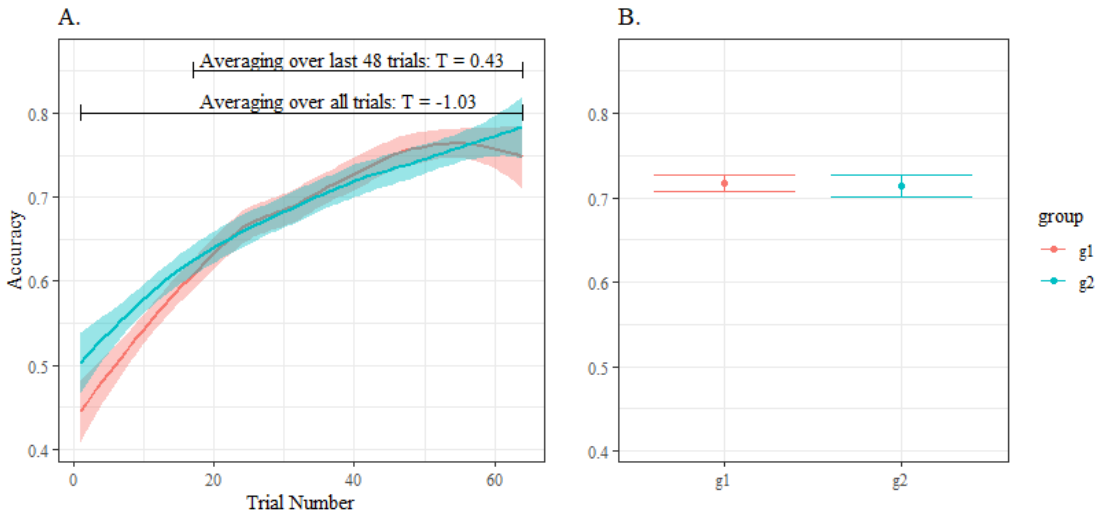


Figure 4. A simulated comparison between groups in which group 1 has lower starting accuracy and faster learning than group 2. No true differences in asymptotic accuracy exist. (A) Lines are LOESS smoothed mean accuracies across all participants within each group, with error bands indicating 95% CI of smoothed means. (B) Means and bootstrapped 95% CI of a standard analysis (i.e., T test comparing final 48 trials).

In this situation the two groups have overlapping levels of accuracy for a period of time. The groups are different in early trials, however. Unlike the previous example, Group 1 would not appear better in any aggregation-based analysis. In contrast, if only early trials were aggregated (i.e., fewer trials collected), Group 2 would appear to have higher accuracy. This would accurately convey group 2's benefit in accuracy start, but it would still be misleading. To the degree that faster learning is considered a good thing, then Group 1 is "better" in one way than group 2. This advantage would not be detectable with aggregation-based analyses.

Discussion

Two simulated experiments showed interactions between generative processes, experimental methods, and analysis choices. Given different true states of the world in the two examples, the inferences supported by T tests comparing group accuracies were highly sensitive to choices of trial number and practice-trial (i.e., discarded data) number. Null, non-null, or even reversed non-null effects may be concluded when aggregating over portions of the available data. Importantly, in each of the specific cases, as well as more generally, aggregation-based analyses are unable to distinguish between the three possible differences in generative process that could have given rise to group differences. The three possible differences in generative process are not simply artifacts of the data simulations here, but instead correspond theoretically relevant and interpretable ways in which groups could vary (for further exploration of these dimensions see Study 5.2). For instance, variation in initial performance is independent of within-task learning and may be constrained by fundamentally different mechanisms than asymptotic performance (i.e., lower-level procedural and pattern learning produces increasingly expert-like performance with extensive experience; Ackerman et al., 1995)

There is no intended claim in these simulations that starting, asymptotic, and learning-rate differences between groups will be appropriate to test in all behavioral data. Quite to the contrary, the intention of these demonstrations has been to encourage a fresh reconsideration of the interactions between theory, methods and analysis. In many situations these three variables may not all be relevant. For example, in a category learning study with 2 novel categories to learn, the methods may be designed such that there is no theoretical reason to believe participants would initially perform differently than 50% correct (see Study 3.2 Experiment 1). It would therefore likely be *a priori* unnecessary and unjustified to test for initial differences in categorization performance (as these would, by definition, arise from chance) and it would, instead, be more theoretically justifiable to set initial performance to be equivalent. In the same study, if the categories were sufficiently discriminable (e.g., with valid and reliable feedback), then any high-functioning adult may achieve 100% accuracy with sufficient practice. Asymptotic performance would thus be an irrelevant parameter to model in this case.

Stated more succinctly, when theory and methods both conspire to justify an analytical assumption like a specific level of performance, that assumption should be made. This is the case when values must be estimated as well as when they can be known *a priori*. This manuscript contains several such combined approaches, such as Study 3.3, in which learning-rate and starting parameters were estimated for both training and generalization of learning while asymptotic performance was estimated as being identical for both learning phases.

As described above, if a strong case can be made for a theory only concerning overlearned expert-level asymptotic performance, analyzing only the last few trials may be justified, although even in this case there is little reason to eschew time-dependent detrending (implemented in Chapter 6). Indeed, testing for the relative improvement in goodness-of-fit metrics with learning parameters may be a central question (see Study 3.1). In many circumstances 3-parameter fits improved information criteria values. Increased numbers of parameters may improve these fits even better (see Study 3.3) but may also come at the cost of model interpretability. In Chapter 6 I describe the statistical package I have written (**TEFits**), in which I make every attempt to retain intuitive interpretability of parameter values. Unfortunately, complex interactions between parameters make that goal intractable at times.

On the other hand, possible sources of underfitting (i.e., using too few parameters to appropriately characterize data) that are likely present in many experimental datasets include fatigue and trial blocking (i.e., warm-up). The latter is due to disjunctions in data collection introduced by methodological constraints, and a preliminary parameterization is available in **TEfits** (see also Adams, 1952; K. M. Newell et al., 2001). Fatigue will not be analyzed in the empirical work presented, but implementations in **TEfits** are forthcoming.

In the next Chapter I turn from simulated data to real data. I present two papers, published by myself and my colleagues, demonstrating the empirical utility of time-evolving models – both in terms of simply capturing the data more appropriately than methods that fail to consider time-dependence and in terms of facilitating inferences regarding learning generalization. I then report the

results of an unpublished study in which I investigated the most appropriate mathematical functions characterizing learning curves.

Chapter 3. Empirical and theoretical reasons for a continuous-time approach to learning

Study 3.1 Theoretical and empirical benefits of continuous-time approaches to learning: Kattner, Cochrane and Green (2017)

Chapter 2 demonstrated several examples where there is divergence in the inferences supported by aggregated versus continuous-time approaches to psychological processes. Differences reach far beyond simplistic simulations as well. Theories of learning, ranging from Hebbian and delta-rule models (Bejjanki et al., 2011; Rosenblatt, 1958; Rumelhart et al., 1986) to Bayesian systems (Michel & Jacobs, 2007), presume changes in knowledge representations with every learning event. The distinction may be circular; learning events are events in which knowledge representations have changed in response to behavior and/or the environment. In contrast to this ubiquitous characterization of the nature of learning, conventional approaches to the analysis of learning have measured performance by grouping learning events into blocks.

As I have already briefly touched upon, categorization of time (e.g., into *first block of 100 trials* and *last block of 100 trials*) violates the core theory of the learning being studied. As such, the first step to developing a learning-centered approach to experimental psychology should involve a direct assessment of the empirical implications of conceptualizing changes in processes as occurring on each learning event. In this comparison the potential benefits of a continuous-time perspective on psychological change should be most evident in domains where learning reliably occurs and where many behavioral trials are available to provide stable convergence of parameters. The timescale, stimulus parameters, and the related psychometric function are all well-established in the field of visual perceptual learning, thereby providing a solid foundation for comparing blocked-time to continuous-time approaches to understanding the time course of learning.

Well-established empirical and theoretical bases led my colleagues and I to situate our first implementation of a continuous-time model of learning in the field of visual perceptual learning (Kattner,

Cochrane, & Green, 2017). We compared time-continuous analytical approaches to standard analytical approaches (i.e., time-evolving summary statistic vs. calculating summary statistics for discrete blocks). Our findings, of consistently improved model fit indices of the continuous-time approach over a blocked-time approach, laid the foundation for substantive work on continuous processes of learning reported later in this manuscript. Kattner, Cochrane, and Green (2017) provided the empirical justification for implementing continuous analyses of learning while also providing the analytical framework for fitting time-evolving psychometric function thresholds (see Appendix 2).

Study 3.2 Continuous-time approach allows novel inferences about processes of generalization: Kattner, Cochrane, Cox, Gorman, and Green (2017)

Kattner, Cochrane, and Green (2017) provided both theoretical and empirical justifications for a continuous-time approach to understanding processes of learning. By statistically modeling psychometric functions as continuous changing throughout learning goodness-of-fit measures were improved and theoretical constraints were satisfied. However, Kattner, Cochrane, and Green (2017), by virtue of being fully focused on examining the ability to better fit data using continuous time as compared to block approaches, did not include a direct demonstration of additional inferences that are possible when treating learning as a continuous process. This instead was the focus of our next published article, where my colleagues and I tested the idea that our inferences regarding learning generalization would be facilitated by a continuous-time approach.

The question of generalization after perceptual training is central to understanding the mechanistic underpinnings of the improved functioning (Ahissar & Hochstein, 2004; Ball & Sekuler, 1987; Dosher et al., 2013; Fahle, 2005; Jeter et al., 2009). Process-level improvements are paradigmatically probed by comparing patterns of generalization between training conditions. Generalization or specificity to novel distributions of stimuli (e.g., orientation angles) or to novel retinal locations have been used, under some theoretical frameworks for instance, as evidence for the putative level of processing at which learning occurred (Ahissar & Hochstein, 1997; Cochrane et al., 2019).

Despite the ubiquity of generalization as a methodological tool to investigate perceptual improvements, possible mechanisms of generalization are obscured or conflated when using measures predicated on stationarity within blocks of generalization trials.

The most frequent interpretation of successful generalization is that learners' training-task improvements provide immediate benefits to novel tasks (e.g., after changing stimulus parameters or retinal location). Immediate benefits in turn connote improvements in shared processes (i.e., processes involved in both training and generalization). Yet any aggregated analysis predicated on *iid* measurements over many trials combines initial performance with later performance, thereby preventing inferences regarding immediate transfer. Instead, other mechanisms of generalization (e.g., improved ability to learn) may lead to identical results in block-level analyses. That is, conventional block-wise analytical techniques cannot provide the information required to conventionally adjudicate between processes of immediate transfer (i.e., learning on Task 1 produces an immediate increased ability to perform Task 2) and enhanced learning (i.e., learning on Task 1 produces an increased ability to learn to perform Task 2).

In order to investigate the differential mechanisms of generalization of perceptual learning we demonstrated that blocking behavioral trials results in two unintended consequences. First, blocked analyses reduce or preclude differentiation between generalization as a process of immediately improved performance versus a process of increased speed of learning. Second, blocking trials carries the theoretical assumption that trials within blocks are *iid* and that change in abilities can only occur between blocks, thereby making the analyst's choice of block size into a consequential decision. As such, we demonstrated that the choice of block size may make generalization significant or non-significant, which is a clearly undesirable researcher degree of freedom (Figure 4, Kattner, Cochrane, Cox, et al., 2017; i.e., this provided a direct empirical demonstration of the issues described in hypothetical data in Chapter 2). A continuous-time approach to interpreting results avoids each of these problems by leveraging all data to understand the full trajectory of learning and generalization (see Appendix 3).

Study 3.3 Continuous-time models provide novel inferences about perceptual learning functions and generalization

Our previous work demonstrated the theoretical and empirical justifications for implementing continuous-time models of learning (Kattner, Cochrane, & Green, 2017). This study used parameterizations of learning (i.e., a simple exponential form) that were motivated by previous comparisons of the mathematical underpinnings of perceptual learning (e.g., Dosher & Lu, 2007; Klein, 2001). However, those previous assessments of the functional forms of learning have largely utilized aggregation across learners and/or across trials (e.g., Dosher & Lu, 2007; c.f. Gallistel et al., 2004; Karl M. Newell et al., 2009). Thus, while the learning functions utilized in the previous two published reports certainly outperformed the aggregate-based analysis alternatives, this does not reflect compelling evidence that this form was the 'right' functional form.

Importantly, understanding the mathematical functions describing learning is not simply an empirical data-fitting problem. Instead, uncovering the best fitting functional form for learning has both purely descriptive as well as deep inferential repercussions. Descriptive understanding of the data through reduction to a mathematical function (e.g., the ubiquitous linear model) provides leverage for inferences about higher-order statistical questions. One prominent example is neuroimaging, in which case parallel generalized linear models may be used to create activation maps, which themselves may be used to draw inferences regarding localization of function. Likewise, in studies of learning, change scores (i.e., pre-test mean minus post-test mean) are often used to contrast learning conditions; these change scores implement an implicit linear model of change linking condition, pre-test scores, and post-test scores.

As should be clear from the examples above, assumptions of the functional forms of learning are implicitly implemented in many common practices. Linear assumptions are by far the most common (e.g., change scores, linear trends, and certain growth curve models). The basic implications of linear models immediately and *a priori* show the problems of this assumption. Linear change implies a constant change in the outcome per unit time, which necessarily means that an arbitrarily large amount of time learning

would lead to an arbitrarily large (positive or negative) level of performance (see Figure 5). While the extrapolation problem may seem like simply a nuisance, it is in fact a fundamental threat to analyses that assume linearity. The assumption of linearity across time constrains learning rates [slopes] due to the high statistical leverage of late performance. Changing the amount of learning time on an identical task (e.g., from 50 trials to 500 trials) thus changes the possible range of estimates of learning rate. Any line fit to 500 trials would be necessarily flatter than a line fit to 50 trials (i.e., smaller rate of learning; see Figure 5). This is in contrast to dominant learning theories; the first 50 trials should involve the same rate of learning regardless of the presence or absence of the following 450 trials (under the assumption of unitary processes of change; see K. M. Newell et al., 2001). Further, within a constant amount of time, individuals' linear learning rate is fully conflated with their magnitude of learning (Primi et al., 2010). Linearity is thus statistically inappropriate for the study of learning.

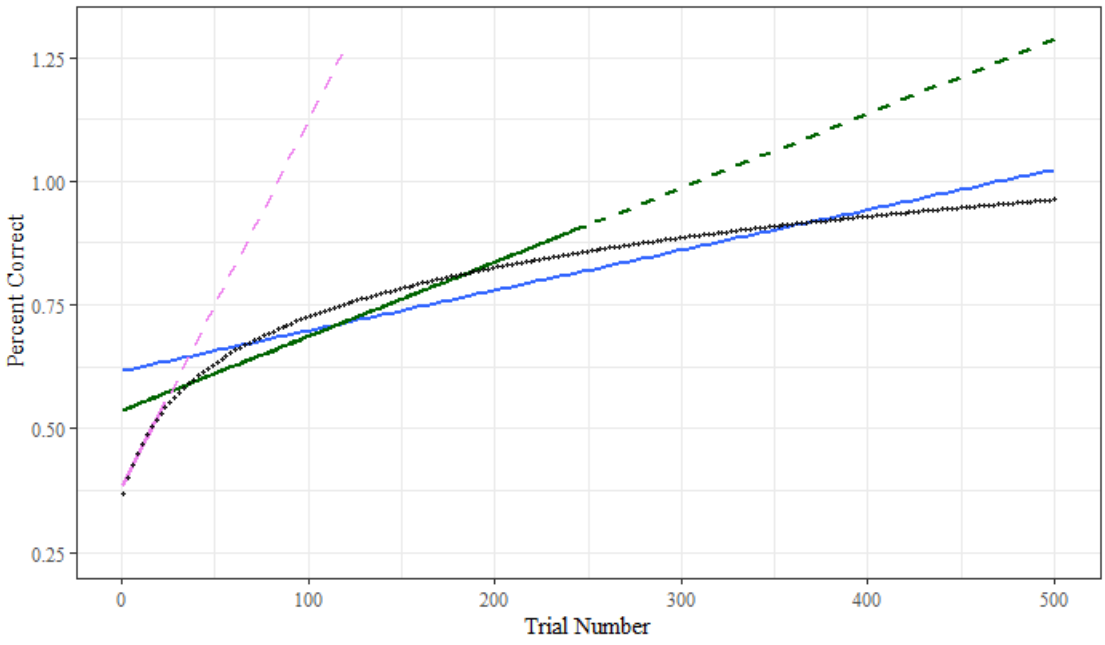


Figure 5. Implications of the timescale-dependent biases of linear fits. Given the same learning curve (black dotted line), linear fits (solid lines with dashed extrapolations) may vary widely. Linear fits to 50 trials (pink), 250 trials (green) or 500 trials (blue) each provide very different inferences regarding

the nature of learning. The 50-trial linear fit necessarily indicates rapid learning that quickly extrapolates to impossibly large values of accuracy. The 500-trial linear model necessarily fits a flatter line (i.e., putatively "slower" change) than the models fit to smaller number of trials. The 500-trial model thereby misses, by design, the early rapid changes of the learning curve. The slopes, beginning levels, and ending levels of performance diverge between all models.

Linearity is likewise empirically inappropriate for learning. Instead, for nearly a century, researchers have characterized learning using linear models using the logarithm of performance (in the case of exponential functions), logarithms of performance and time (in the case of power functions), or other empirically appropriate (i.e., providing good fits to data) functions (Snoddy, 1926; Wright, 1936). In essentially all cases, the functions that have been employed are saturating, that is, an extrapolation of the function to an arbitrarily large amount of time causes the slope of the curve to converge on zero. Indeed, a slope approaching zero with increasing learning time is a necessary feature of all time-dependent models considered in this manuscript (for other approaches see Deboeck et al., 2009). Further, in classic texts using linear functions of log-transformed data, these curves necessarily converge on point estimates of zero (or some other predetermined constant; Karl M. Newell et al., 2009).

By far the most dominantly adopted learning form is the power function (e.g., Anderson et al., 1999; A. Newell & Rosenbloom, 1981; Snoddy, 1926; Stratton et al., 2007). This is due in large part to an influential review and comparison of learning functions providing support for the so-called "power law of practice" (A. Newell & Rosenbloom, 1981). In this review of the learning literature, linear analyses of transformed data were used to compare learning functions, largely using group-level response time data from several classic motor learning studies. The authors concluded that the power function was applicable to learning in domains such as perceptual, motor control, decision-making, problem solving, and memory. The dominance of the power functions was entrenched by its use as a default implementation of learning in many settings, such as formal systems like ACT-R (Anderson et al., 1999). Nonetheless, power functions' appropriateness as an empirical description of learning has been called into question by

researchers in motor learning, perceptual learning, and similar fields (Dosher & Lu, 2007; Heathcote et al., 2000; Stratton et al., 2007).

Beyond improvements to data understanding through accurate description and reduction, underlying mathematical functions also allow for inferences regarding processes giving rise to observed behaviors. Exponential decay, in which the proportion of change remaining is a function of time elapsed, implies a single mechanism of change (i.e., each unit of input [time] produces one particular proportional amount of change). Power-law functions, in contrast, imply multiple mechanisms of change; just as an exponential function implements a single rate of change per unit of input [time], a power function implements a slowing rate of change per unit [time]. Exponential and power functions each can be parameterized with 3 parameters, with these simplest parameterizations forming the basis of the learning family (e.g., Heathcote et al., 2000).

Augmentations to simple learning parameterizations allow for implementation of specific additional hypotheses. In the case of power functions, one common example is an additional parameter to model theoretical "quantity of prior learning" (Heathcote et al., 2000). While adding a great deal of flexibility to the possible shapes of power functions, this additional parameter also carries a clear and interpretable meaning. Additions to the exponential function allow, for example, modeling of multiple learning processes using several rate parameters (Karl M. Newell et al., 2009; Reddy et al., 2018) or an initial acceleration or "slow start" of learning (Brooks et al., 1995; Leibowitz et al., 2010; K. M. Newell et al., 2001). The cumulative Weibull function adds a single shape parameter to the 3-parameter exponential function and thereby allows for an interpolation between the simplest exponential function and a fully sigmoid function akin to a logit or probit (Gallistel et al., 2004). The shape parameter corresponds to a deceleration or acceleration of the hazard rate (i.e., proportional learning on each trial), with a more sigmoid-like function indicating a slow start to learning and an accelerating hazard rate.

As noted above, in my previous work I have utilized the simplest exponential form to model learning. However, I did not have the necessary dataset to adjudicate between that and other functional forms that can take on similar (but not identical) shapes. As such, here I took inspiration from the

perceptual learning studies of Ahissar and Hochstein (2000) and Wang and colleagues (2013), using abbreviated methods for each (i.e., 2 days total of learning). Each of these previous studies demonstrated learning in visual perception as well as generalization of learning modulated by task difficulty (see also Ahissar et al., 2009). In the first set of Results and Discussion I therefore compare goodness-of-fits across the candidate functional forms described above. The best fitting model (which to preface the results, differed across tasks), thus provides information regarding the learning processes involved in these two tasks. Beyond examining the functional form of learning, this study also had a second purpose – which was to examine learning generalization in a time-dependent fashion. Indeed, generalization in the field of perceptual learning is nearly always tested via a single block of a new task with performance on the new task being aggregated over the entire block (e.g., trained for 1000 trials on Task #1, and then generalization assessed via 100 trials on Task #2). In the second set of Results and Discussion I use each experiment's best-fitting by-trial functions of learning to assess the extent to which generalization of learning was modulated by task difficulty, and whether the basis of generalization was due to changes in initial performance or learning to learn (see Study 3.2). Both purposes though were motivated by the necessarily by-trial nature of learning, and the corresponding increase in theoretical insights provided by analyses that conform to the by-trial nature of learning.

Methods

I recruited participants from the University of Wisconsin–Madison Introduction to Psychology participant pool. All participants read and signed consent forms and were compensated with course credit. All procedures were approved by the University of Wisconsin–Madison Institutional Review Board.

Participants were assigned to one four groups. Two groups completed a texture oddball detection task (Ahissar & Hochstein, 1993, 2000; see Figure 6.A.) while two groups completed a dot-motion direction discrimination task (Ball & Sekuler, 1987; Liu, 1999; X. Wang et al., 2013; see Figure 6.B.). Within each of these tasks, one group completed an easier version while one group completed a more difficult version. Participants were assigned pseudorandomly to difficulty group.

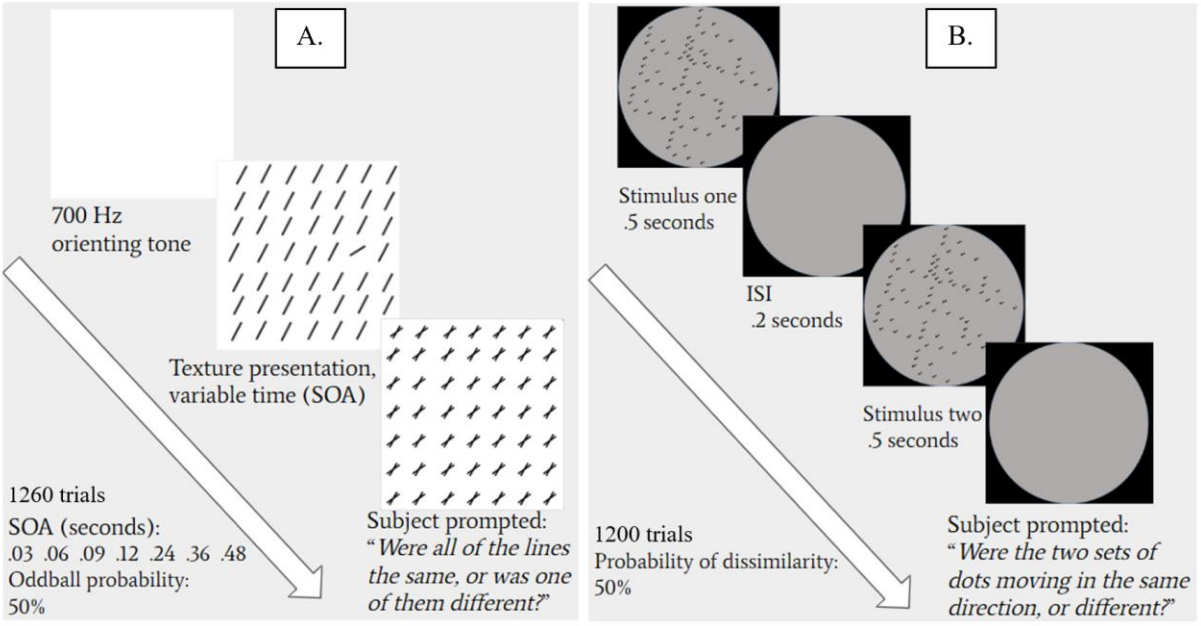


Figure 6. Depictions of each perceptual learning paradigm. (A) Texture oddball detection task and (B) Dot-motion direction discrimination task. For further details see Appendix 4.

Procedure

All participants completed two days of training involving 8 blocks of a perceptual learning. The first six blocks, four on the first day and two on the second, involved training on a single stimulus reference orientation (i.e., texture orientation or mean dot-motion direction). The last two blocks tested generalization of learning by training on an orthogonal reference orientation. Before the first block and the seventh block, each participant completed 4 trials with very large orientation offsets and slower timing, in order to familiarize participants with the orientations within the phase of the task. Training stimulus sets remained constant throughout the experiment, allowing performance changes to be unbiased by inter-participant variation in trial histories. This contrasts with, for example, staircase methods which conflate stimulus difficulty with participant ability.

Two perceptual learning tasks, designed to replicate Ahissar and Hochstein (2000) and Wang, Zhou, and Liu (2013; only within-difficulty generalization) were employed. Task parameters were largely

taken from these studies (see also Figure 6). Two difficulty orientation offsets were used in each task: easy (texture: 30°; motion: 8°) and difficult (texture: 16°; motion: 4°). All participants were trained for approximately 1 hour on each of 2 days (840 per day texture, 800 per day motion). The first 75% of trials used one reference angle (16° texture, 40° motion), while the last 25% of trials tested for generalization to another reference angle (106° texture, 130° motion). Additional task information is reported in Appendix 4.

Analysis

Nonlinear learning models were fit in R using the **TEfits** package (model code reported in Appendix 4). As in the original studies cited above, outcomes in texture detection were defined as thresholds while outcomes in dot-motion were defined as d-prime. Learning functions are described in detail in the **TEfits** documentation. In Ahissar and Hochstein (2000) texture detection thresholds were fit using "Quick function," an alternative parameterization of the Weibull function, whereas **TEfits** uses a different Weibull parameterization of the psychometric function (see **TEfits** documentation).

Analyses were guided by the goal of identifying the best-fitting mathematical function describing changes in performance associated with practice. This necessitates fitting parameters at the levels of individual participants (Dosher & Lu, 2007; Heathcote et al., 2000; Stratton et al., 2007) and using time as a continuously varying dimension (Kattner, Cochrane, & Green, 2017). Nonlinear regression is necessary in applications wherein the relationship between predictors and outcomes cannot be coerced into a linear function (see Chapter 6 for discussion). Within participants, models simultaneously estimated parameters for initial learning and subsequent generalization.

The primary candidate functional forms of learning included 3-parameter power and exponential functions, as examined in similar previous studies (Crossman, 1959; Dosher & Lu, 2007; Heathcote et al., 2000; Leibowitz et al., 2010; A. Newell & Rosenbloom, 1981; Karl M. Newell et al., 2009; Snoddy, 1926). Broadly, the exponential family takes the form $[start + (asymptote - start)^{time*rate}]$ while the power family takes the form $[start + (asymptote - start) * time^{rate}]$. Each of these functions allows a parameterization with three free parameters describing (1) the starting point of performance (i.e.,

y-intercept), (2) the rate or shape of change, and (3) the asymptotic level of performance expected with an infinite amount of experience. A common augmentation of the power function includes an extra parameter, conceptualized as “amount of previous experience” (Heathcote et al., 2000), which adds a great deal of flexibility to the shape of the power function. Two four-parameter extensions of the exponential function were also tested: (1) a weighted combination of two exponential functions (Reddy et al., 2018), and (2) a Weibull function (Gallistel et al., 2004). The Weibull function is of particular interest; it is an extension of the exponential function to learning that may start slowly and form a sigmoid learning function.

In each case, all possible models of interest were fit to both initial training and test of generalization. This included common parameters between training and generalization *except* for initial ability and rate of learning. Start and rate parameters were estimated as varying between training and generalization, thereby allowing for tests of generalization. Representative learning curves for each function can be observed in the Results. Model comparisons were conducted by first normalizing model BIC within participants to extract Schwartz weights (Wagenmakers & Farrell, 2004). Schwartz weights must sum to one within participants, with the highest weight indicating the best-fitting model.

Results: Data processing and exclusions

The texture oddball-detection task started with 26 easy-condition participants and 30 hard-condition participants. 6 easy- condition participants and 11 hard-condition participants were excluded due to (1) not achieving above-chance performance on their final 200 training trials (i.e., failing to reject the null of a one-tailed binomial test) or having accuracy on the final 200 trials of training that was lower than accuracy on the first 200 trials of training. This left 20 easy-condition participants and 19 hard-condition participants. The dot-motion perceptual learning started with 31 easy-condition participants and 45 hard- condition participants. 8 easy-condition participants and 25 hard-condition participants were excluded due to (A) not achieving above-chance performance on their final 200 training trials or (B) having accuracy on the final 200 trials of training that was lower than accuracy on the first 200 trials of training. This left 23 easy-condition participants and 20 hard-condition participants. The rate of

exclusions was quite high, particularly in the dot-motion perceptual learning. The high number of exclusions was unfortunate, but the current study of learning necessitated a confirmation that all participants learned. Only if participants learned could parameterizations of the learning function be assessed.

Figures 7 and 8 show the fit values of each learning function to data from an example participant from each study. Characteristic differences between the power family and exponential family were clear in the shapes of the learning curves during training, while the learning curves during generalization were temporally limited enough that family differences were less clear. The Weibull function's potential for a slow start of learning was pronounced in the dot-motion learning paradigm but not the oddball-detection learning paradigm.

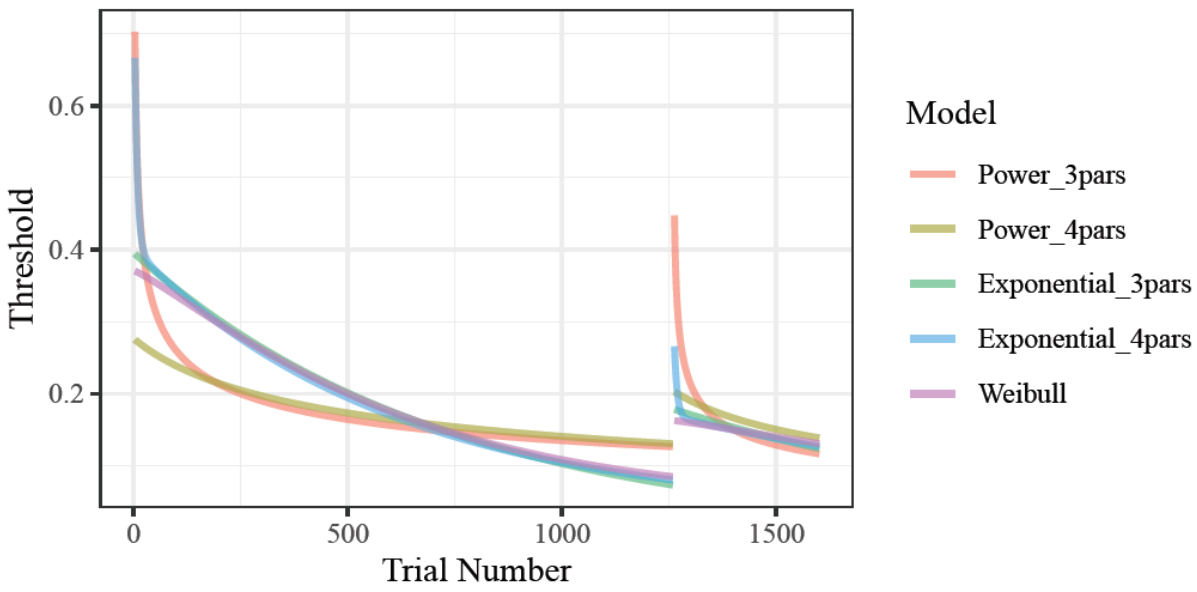


Figure 7. Predicted time-evolving psychometric function threshold from each learning function fit to participant S191972. The x axis is overall trial number including both initial training and generalization (the disjunction in the curves occurs on the trial where participants switch from initial training to the generalization task). The y axis is Weibull psychometric function threshold indicating the number of

seconds needed to achieve 75% accuracy. Lower threshold values indicate better performance. Learning curves display typical differences between the power family (Power_3pars and Power_4pars) and the exponential family (Exponential_3pars, Exponential_4pars, and Weibull).

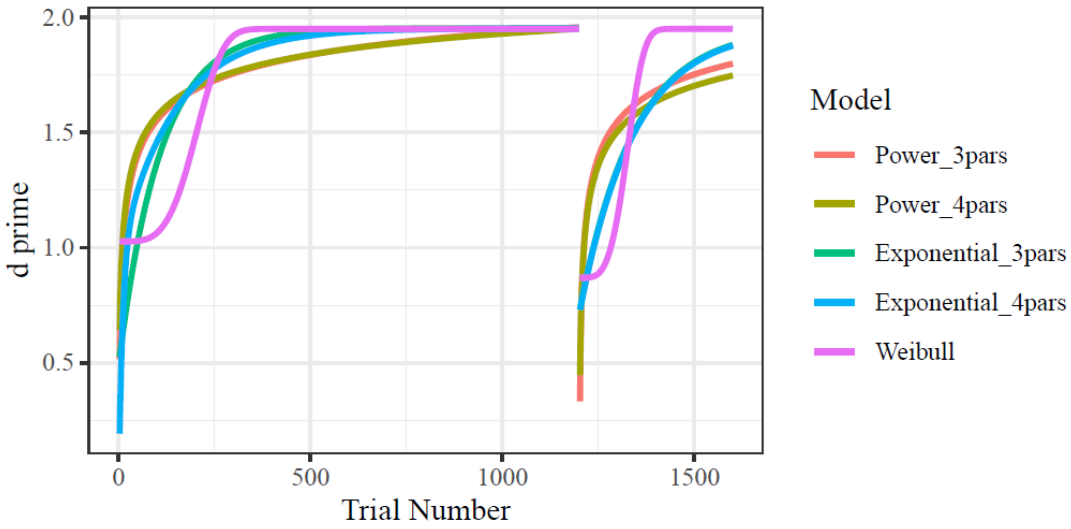


Figure 8. Predicted time-evolving d prime from each learning function fit to participant S191920. The x axis is trial number, with initial training and subsequent generalization modeled together. The y axis is d prime, or sensitivity, with greater d prime indicating superior performance. The power family functions appear similar to one another, as do the two exponential functions. The Weibull function is strikingly different than the other four functions, however, by allowing for an initial acceleration of learning and a sigmoid shape.

Results: Functional forms of learning

The primary analysis of interest involved comparisons of the relative evidence for each model (i.e., Schwartz weights). The simplest function from the exponential family, the 3-parameter exponential, was the best-fitting model to the vast majority of texture oddball-detection participants (see Figure 9). The weights of the 3-parameter exponential did not vary by difficulty (easy best fit = 95% of participants;

difficult best fit = 89.5% of participants; coefficient of difficulty predicting 3-parameter exponential weights *b* = -0.023, CI = [-0.11,0.069], dR^2_{oos} = -0.011).

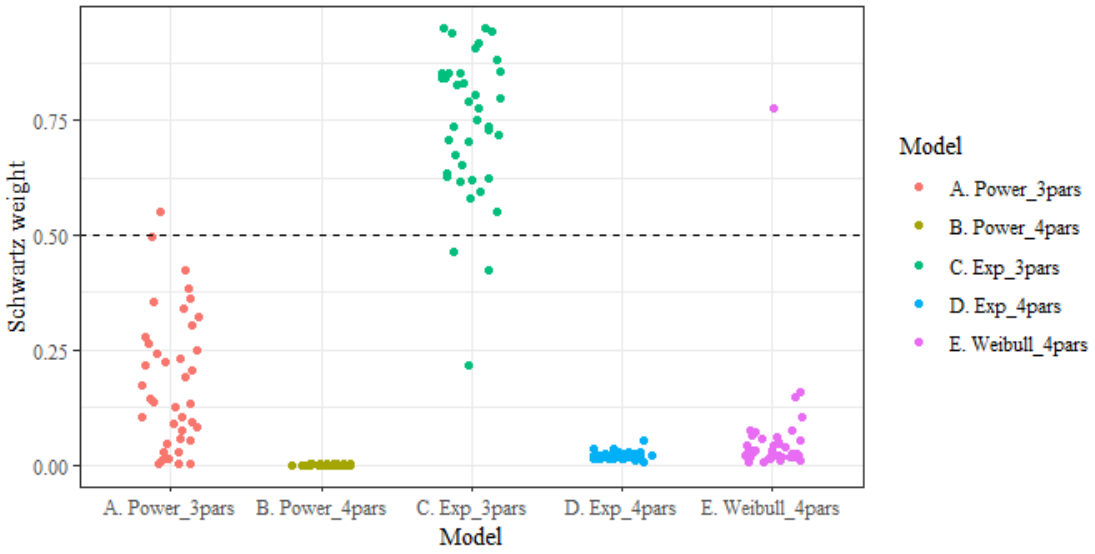


Figure 9. Texture oddball detection relative evidence for models fit using each learning function. Each participant has a weight (point) for each model. Weights are normalized BIC within participants and across models. All points above the dashed line indicate that, for a given subject, the model received more weight than all other models combined.

Learning in dot motion direction discrimination tended to be best fit by the Weibull function (see Figure 10). The weights of the Weibull models did not vary by difficulty (easy best fit = 69.6% of participants; difficult best fit = 60% of participants; coefficient of difficulty predicting Weibull weights *b* = -0.06, CI = [-0.42,0.2], dR^2_{oos} = -0.0142). Of the participants whose learning was best characterized by the Weibull function, 93.1% had shape parameters over 0, indicating the benefits of the extra parameter were derived from fitting learning with a slow start (i.e., sigmoid shape and increasing hazard function).

Nearly all of participants who were not best fit by the Weibull function were instead best fit by the 3-parameter exponential function (25.6% of all participants).

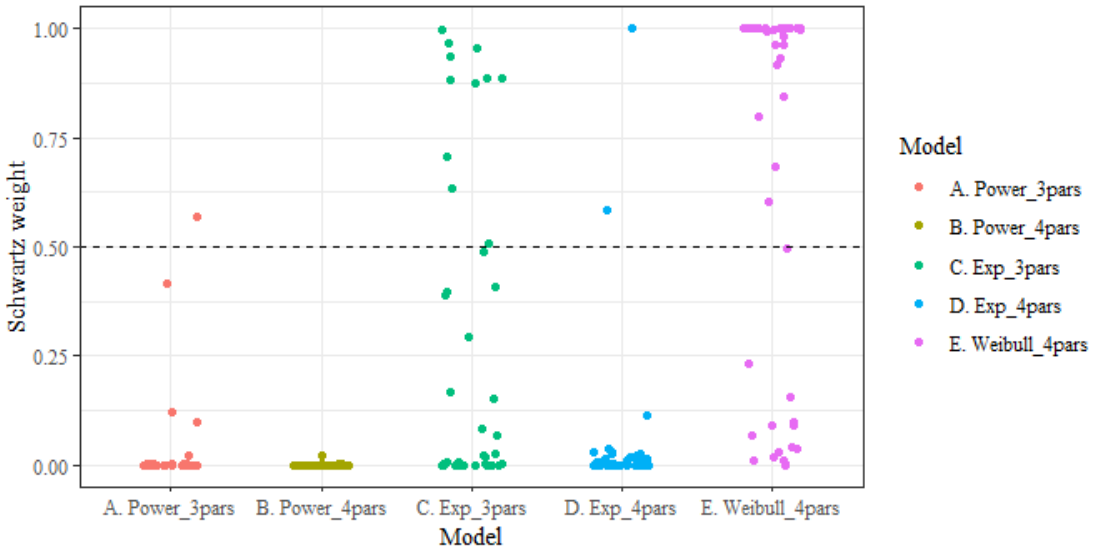


Figure 10. Dot-motion direction discrimination relative evidence for models fit using each learning function. Relative evidence for models fit using each learning function. Each participant has a weight (point) for each model. Weights are normalized BIC within participants and across models. All points above the dashed line indicate that, for a given subject, the model received more weight than all other models combined.

Discussion: Functional forms of learning

In two perceptual learning experiments, novel comparisons between functional forms of learning were implemented on a by-trial and by-participant scale. Functions coming from the exponential family provided the best trial-level fits to almost all participants' learning. In an oddball detection task using oriented-line textures improvements in psychometric function threshold were overwhelmingly best characterized by a simple 3-parameter exponential model of change. Within a dot-motion direction discrimination task, improvements in d prime were best characterized by a Weibull or a 3-parameter

exponential function. Because the 3-parameter exponential function is nested within the Weibull function (i.e., the latter is an extension of the former with one extra parameter), both experiments repudiate the "power law of learning" (Newell & Rosenbloom, 1981) and corroborate evidence for exponential functions over power functions (Dosher & Lu, 2007). The dominance of the exponential family was uniform across levels of task difficulty, thereby precluding the possibility that difficulty-related variations in learning may lead to distinct trajectories (M. Ahissar & Hochstein, 1997).

The nested nature of the 3-parameter exponential function within the Weibull function carries the implication that all models utilizing the 3-parameter exponential function would be equally well fit by the Weibull function. That is, the 3-parameter exponential function is a special case of the Weibull function, and as such all learning best fit by the 3-parameter exponential function can likewise be seen as being characterized by a special case of the Weibull. Model comparisons between these two models served simply to test whether the extra "start speed" (i.e., shape) parameter of the Weibull function was statistically justified through a sufficient reduction in error. The results of the dot-motion learning experiment demonstrate the utility of comparing these nested models to assess the appropriateness of the shape parameter. Although all learning was best fit by a Weibull function, some of these Weibull functions could be reduced to the 3-parameter exponential function without compromising model fit (Gallistel et al., 2004).

Support for an exponential family of change implicates a mechanistically simple process of learning in each paradigm. The novelty of the continuous-time approach to learning thereby corroborated earlier work in perceptual learning which supported a unitary process of change underlying learning (Dosher & Lu, 2007). While oddball detection learning was best characterized as arising from learning a constant proportion on each learning event (i.e., constant hazard rate), in many participants dot-motion direction learning was better characterized by including a parameter that allowed the hazard rate to either increase or decrease. The overwhelming majority of these participants displayed an acceleration in learning (i.e., a sigmoid function). An increasing hazard rate appears to indicate an interaction between a simple exponential mechanism of learning and a second mechanism that prevented immediately constant

learning and thereby slowed down initial learning. This mechanistic explanation remains speculative, however, and the current results provide the impetus for a more thorough investigation of the potential dissociations or interactions between simple learning processes and modulatory mechanisms that may inhibit the initial learning rate. It may be possible to identify distinct processes, for example, by using convergent methods such as concurrent measurements of behavior, movement kinetics, and event-related potentials (Gratton et al., 1992). In addition, future work should explore the experimental conditions and the performance indices for which initial learning is slowed down in comparison to conditions under which the hazard rate is constant.

Difficulty modulations of generalization

As noted in the introduction, the present study was designed both to assess the functional form of learning and to utilize novel by-trial models to test the claims for difficulty-modulated generalization in perceptual learning. In previous research utilizing both experimental paradigms the degree of observed learning generalization has been found to be modulated by the difficulty of the training task (Ahissar & Hochstein, 1997; Liu, 1999). Specifically, greater learning generalization has been seen when participants were trained on easier, as compared to harder versions of the tasks. The enhanced generalization in response to easier training regimes was then in turn seen as evidence for distinct mechanistic loci of learning, with mechanisms of learning on easy tasks being more general and mechanisms of learning on difficult tasks being more specific (reviewed in Ahissar et al., 2009).

Each of these claims relied on aggregated measures of performance, thereby possibly conflating processes of generalization that may have been time-dependent (i.e., initial performance vs. learning to tune visual perception to novel stimuli; see Ahissar et al., 2009, Figure 4). Although some consideration of time-evolving aspects of learning have been considered, such as linear models fit to block-level d prime (Wang et al., 2013), no previous examination of generalization has used functionally appropriate methods sensitive to rapid changes in performance. By using by-trial models of the most empirically appropriate functional I specifically tested for two possible routes to difficulty modulations in generalization. First, I tested whether immediate benefits of previous training would be evident at the start

of generalization. Second, I tested whether learning within the generalization task would be faster or slower as a function of training task.

Results: Generalization by difficulty

Given the lack of differences in functional form by difficulty, using each experiment's most common best-fit models (i.e., 3-parameter exponential or Weibull) I next assessed the degree to which training task difficulty influenced generalization. Generalization may have happened either immediately or via changed rate of learning. I first tested the possibility of difficulty-related generalization via changes in initial performance on orthogonal stimuli.

Generalization parameters were fit as offsets from the training-period parameters. As such, dot-motion starting generalization parameters above zero or oddball-detection starting generalization parameters below zero would indicate some generalization of learning. Figure 11 shows that easy conditions did appear to lead to greater generalization of learning in the form of starting ability.

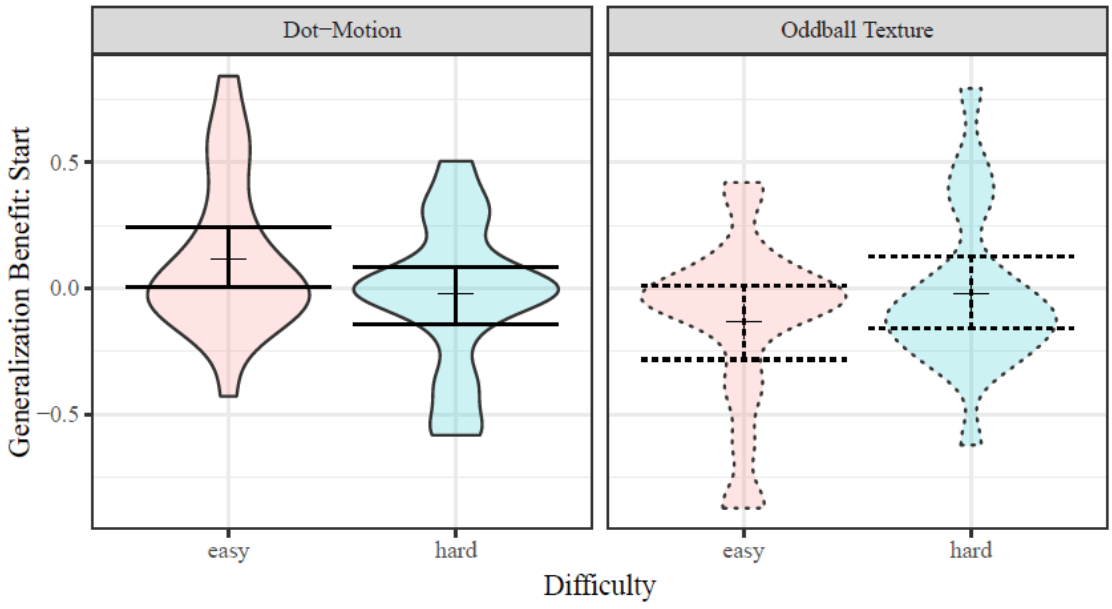


Figure 11. Differences between generalization start and training start. Larger values indicate beneficial generalization for Dot-motion starting d prime while smaller values indicate beneficial generalization for Oddball Texture starting thresholds. Generalization appears zero-centered (indicating specificity) in each

hard condition, while each easy condition demonstrates an offset toward beneficial generalization (toward higher values in the case of the dot motion task and lower values for the texture task). Dot-motion parameters were divided by 5 to be on a similar scale as Oddball Texture. Shaded area indicates smoothed density. Lines indicate mean and bootstrapped 95% CI.

In each task I tested for the effect of difficulty on starting generalization in two ways. In the first method I fit a bootstrapped robust linear model predicting the generalization starting parameter using difficulty condition while controlling for each participants' training start parameters. In the second I normalized the magnitude of the generalization parameter by dividing it by the difference between a participant's asymptotic level and their training start. I then fit a bootstrapped robust linear model to predict this proportional benefit of generalization using difficulty condition.

In oddball texture detection, training on the easy condition generalized more than that training on the difficult condition (b = 0.11, CI = [0.01,0.2], dR^2_{oos} = 0.0974). However, this effect was lost when the generalization benefit was calculated as a proportion of total learning (b = -0.12, CI = [-0.86,0.62], dR^2_{oos} = -0.0146). Likewise, when examining dot-motion orientation discrimination parameters directly, easy condition generalized more than difficult (b = -1.1, CI = [-2,-0.33], dR^2_{oos} = 0.2115). When calculating dot-motion generalization benefit as a proportion of learning there was no difference between difficulties (b = 0.03, CI = [-0.84,0.71], dR^2_{oos} = -0.014). In total these results indicate that difficulty-related patterns in generalization were seemingly driven by differences in the overall magnitude of learning between difficulties.

I next used bootstrapped robust linear models to predict generalization rate using difficulty condition, while controlling for training rate. There were no differences in generalization of learning rate between difficulty conditions for either learning paradigm (see Figure 12).

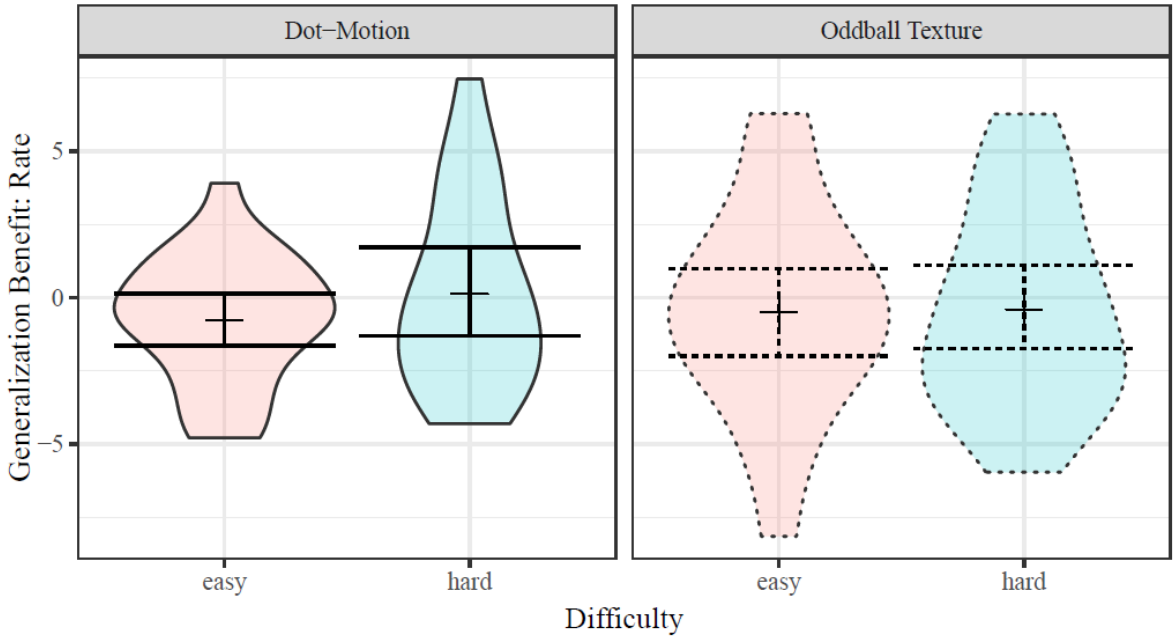


Figure 12. Differences between generalization rate and training rate. Smaller values indicate beneficial generalization in both Dot-Motion and Oddball Texture tasks. Only the easy Dot-Motion condition indicates any beneficial generalization, although the interaction with difficulty is not reliable. Shaded area indicates smoothed density. Lines indicate mean and bootstrapped 95% CI.

Clearly, to the extent that differences in generalization due to difficulty were evident, these group differences were due to immediate benefits in performance rather than differences in the rate of generalization.

Discussion: Generalization by difficulty

After comparisons of learning functions, by-trial models also utilized to modeling to test previously-reported results regarding the influence of task difficulty on the generalization of perceptual learning (Ahissar & Hochstein, 1997, 2000; Ball & Sekuler, 1987; X. Wang et al., 2013). Consistent with previous research, greater generalization was found in both tasks' easy conditions relative to the associated difficult conditions. Critically, though, continuous-time models provided support that this difference was in terms of immediate transfer (i.e., there were differences in the starting points of learning on the generalization task). This approach also uncovered the fact that the effect was not robust when

generalization was calculated as a relative rather than absolute benefit, indicating that overall learnability differences between easy and difficult tasks may determine performance in ways that may imitate disproportionate benefits of easy-task training.

To make this last point more concrete using dot-motion perceptual learning as an example, cross-participant d prime increased from start to asymptote. Although the easy condition's first generalization trials improved when compared directly to their initial performance, the difficult condition's generalization was reliably smaller. However, the fact that the *overall amount* of learning in the easy condition was greater than the difficult conditions means that the definition of generalization itself is not clear *a priori*. The benefits of the easy condition relative to the difficult condition were clear when considering only starting parameters (keeping in mind that the two curves shared an asymptote), yet there was no such effect when considering the overall amount learned in the different difficulty conditions.

The present study loosely replicated two perceptual learning experiments. Continuous-time modeling provided uniform support for unitary exponential learning processes, with a possible need to account for a second process modulating the initial speed of learning (possibly related releasing the "brakes" on learning; Bavelier et al., 2010). There was no support for the use of power functions to understand learning processes, leading to the unequivocal recommendation to use the 3-parameter exponential function or the more flexible Weibull function in order to most appropriately characterize the changes in processes associated with learning. The extent to which these two exponential-family approaches should both be fit, and model comparison used to either justify or reject the extra parameter associated with the Weibull function, will be constrained by the quantity and quality of behavioral data. In experiments with sparse or noisy data it is likely that the flexibility of the Weibull function will lead to overfitting. The evidence in favor of the exponential family across behavioral paradigms and task difficulties, and when incorporating both initial training and later generalization, provides an empirical foundation for the use of this family. In subsequent chapters I will overwhelmingly use the 3-parameter exponential function to model changes in psychological processes over time.

Chapter 4. Continuous-time psychology extends beyond studies of learning: Applications to response competition paradigms

Each study in Chapter 3 applied a continuous-time perspective to learning-related changes within domains explicitly concerned with learning itself. Visual perceptual learning was repeatedly used as a test case for applications of continuous-time approaches to changes in psychological processes. It may not be surprising, then, that an empirical approach was supported that better aligned with base theory than standard approaches (i.e., learning on every learning event rather than enforcing stationarity within blocks). However, the implications of a learning-centered understanding of psychological processes reach far beyond putative studies of "learning itself". As indicated by the simulations in Chapter 2, divergent implications of aggregated or continuous-time approaches to behavior may be evident on timescales more typical of experimental psychology (e.g., fewer than 100 trials, as opposed to the hundreds or thousands of trials involved in perceptual learning experiments).

In this chapter I report the results of several studies in which, rather than utilizing the conventional time-invariant approach to assessing participants' behavior, I leverage the within-experiment dynamics to facilitate more specific process-level inferences. I use two well-established behavioral paradigms, with seminal papers establishing their respective uses each having well over 5,000 citations (Eriksen & Eriksen, 1974; Greenwald et al., 1998). Each of the behavioral paradigms conventionally utilizes measures of performance aggregated across dozens or hundreds of trials' responses and thereby implicitly posits stationary generative processes for each response. By using widely recognized tasks from both cognitive and social experimental psychology I intend to demonstrate the breadth of relevant contexts for a learning-centered approach to experimental psychology. Learning is not a side effect or a nuisance in these contexts; learning is an integral component of the very processes of theoretical interest.

The first two of these experiments involve learning of task statistics within a simple attentional paradigm, the Eriksen flanker task. The last of these studies re-analyzes several sets of Implicit Association Test (IAT) data to uncover learning-related patterns of behavior that would otherwise be

aggregated into a single overall score. In Chapter 5 there is also an example of the same principle, that of leveraging time-dependent dynamics within paradigms typically treated as time-invariant. That study is placed in the following chapter due to the study's emphasis on individual differences.

Study 4.1 Continuous change and statistical learning in response competition: Cochrane et al. (2018)

I first tested the theoretical importance of a continuous-time approach to understanding the processes giving rise to a canonical effect in cognitive psychology, the response (in)compatibility effect (i.e., response competition). One key feature of response incompatibility effects, such as responding slower to a central left-pointing arrow when surrounded by several right-pointing arrows than when surrounded by left-pointing arrows, is the putatively automatic nature of stimulus-response associations. In short, most theories in the field posit that response competition between stimuli associated with automatic opposing actions (i.e., left key versus right key) leads to response time asymmetries between homogenous and heterogeneous displays. Between-person variations in the magnitude of these asymmetries (i.e., aggregated differences in RT between trial types) are then in turn frequently treated as individual differences in basic selective attention processes (Fan et al., 2002; Machizawa & Driver, 2011; Westlye et al., 2011), with related inferences regarding such between-person variation being important to prominent theories of Executive Functions (Diamond, 2013; Miyake, 2000).

Aggregated differences between trial types connotes stationary distributions of response times. In contrast, there is a high probability of learning task dimensions including motor demands, stimulus and cue timing, and the probability of various trial types. Learning trial type probabilities is particularly important in the context of response competition tasks, and reports of altered response compatibility effects in response to biased task statistics have existed for decades (Gratton et al., 1992; Lehle & Hübner, 2008). In the absence of explicit instruction regarding task statistics participants must learn the relevant distributions through experience in the task. Despite the robustness of this effect, no previous work had examined the trial-wise development of statistical learning within response competition. In a standard flanker task, I manipulated the statistics of trial types to alter participants' task learning. I then used by-

trial nonlinear mixed-effects Bayesian regression to fit the evolution of the response compatibility effect over the course of experimental blocks. I found that response time differences (i.e., response compatibility effect) decreased over time (i.e., learning improved performance). Further, task statistics reliably altered the amount of learning. I presented this work at the Annual meeting of the Cognitive Science Society (Cochrane et al., 2018) (see Appendix 5).

Study 4.2 Learning vs. meta-learning in response competition

In Cochrane et al. (2018) I demonstrated that attention could be approached as being continuously modulated in response to task statistics, thereby aligning analyses and inferences with the theoretical position that statistics are learned gradually with an accumulation of task experience. As noted by Cochrane and colleagues (2018), one important question left unanswered was the possibility of learning on timescales beyond task blocks with internally stable task statistics. In particular, there remained the possibility that learning occurred both within blocks (i.e., to statistics of compatible and compatible stimuli) as well as between blocks (i.e., to the possible ranges of task statistics as well as the timing and motor constraints of the task).

To test the possibility of nested timescales I implemented a variation on the methods of Cochrane et al. (2018). Participants completed two sets of flanker task trials, each with a different set of proportions compatible and incompatible. I thereby was able to test the extent to which response compatibility effects were primarily influenced by shorter (within-block) or longer (between-block) statistical learning. However, unlike Study 4.1, by-trial analyses did not indicate systematic changes in response competition effect due to learning. Further, across conditions with or without feedback, statistical learning was not evident when utilizing blocks of 50 trials to quantify changes in response times over time. Instead it appeared as though modulations of attention were occurring on rapid timescales that were too short to be captured by either trial-wise or block-wise analytical approaches. This work was submitted to Attention, Perception, and Psychophysics in early summer 2020 and a “revise and resubmit” recommendation was returned in August 2020 (see Appendix 6).

Studies 4.3(a-c) Learning within the Implicit Association Test

Studies 4.1 and 4.2 showed that within a common response competition paradigm, the flanker task, systematic changes in performance due to learning may have oft-overlooked consequences for the subtracted measures of interest (see also Gratton et al., 1992; Lehle & Hübner, 2008). Despite the putatively low-level conflict resolution interacting to produce trial-type differences in the flanker task, learning within response competition tasks is not a phenomenon only relevant to the fields of attention or executive functions. Rather, response competition is a general method utilized in various fields of psychology to probe mental processes. One prominent example of a response competition task with wide-ranging implications is the Implicit Association Test (IAT). The IAT is a subtractive measure indexing differences in response times between combinations of image and word stimuli (Greenwald et al., 1998). The classic example of the task is measuring response times to categorize words as good or bad when the words are paired with either black or white faces, with the task including all combinations of these stimulus categories. Faster responses to White+Good stimuli compared to Black+Good stimuli indicate a response competition analogous to the flanker task response competition. Response time differences (i.e., compatibility effects) are then interpreted as indicators of implicit bias.

The IAT has been extended beyond studies of the four stimulus categories described above, including face-image dimensions such as age or gender and paired-word dimensions such as competence or leadership. Each of these types of IAT tasks relies on response competition in the same way that the flanker task does; there is an assumption that participants have stationary associations between stimulus dimensions (e.g., Black-White and Good-Bad) and that these stationary associations lead to stable differences in distributions of response times for different combinations of stimuli. As with the flanker task, this assumption may be questioned by recognizing the ubiquity of dynamic learning about properties of the environment such as stimulus sets or statistical regularities.

Beyond the low-level statistical learning of task demands, there are other reasons to believe that the IAT may involve a measurable change in behavior with increasing task experience. While not learning in a traditional sense, alteration of one's behaviors due to social desirability is a core part of typical human social behaviors (e.g., in the context of bias, Plant & Devine, 1998). Any such alteration must

necessarily be predicated on an initial recognition, explicitly or implicitly, of the relevant social context and personal behaviors. That is to say, regulation of one's actions causes a differentiation between overt and internal biases, and this regulation unfolds dynamically within any given context. The differentiation between behaviors and internal biases was a motivation behind the original development of the IAT, but the dynamic nature of behavioral regulation means that it may be occurring even within the context of the IAT. As participants are responding to stimuli they are learning about the world and they may be (implicitly or explicitly) adjusting their behavior.

Individual differences in dynamic suppression of bias within the IAT could lead to several possible patterns of performance. If a person is relatively impervious to the categorical differences being contrasted, they would be expected to *start* with a relatively low bias contrast (i.e., small category difference). If, however, a person was able to suppress their bias very well, then that person would be expected to have a relatively low bias contrast at the *end* of their IAT. If a person was able to rapidly respond to the context and change their behaviors appropriately, that person's *rate* of behavioral change would indicate faster suppression. Notably, each of these three sources of variation in individual differences could be the reason for overall (e.g., average) differences in categories' RT distributions (see Chapter 2).

While the patterns described above may be interpreted as indicating levels of bias or regulation thereof, it is important to note that changes in behavior need not be intentional or even conscious. Likewise, decreases in bias scores may simply be task learning (e.g., increased narrowing of attentional focus to only the word or face stimulus on a given trial). The results reported here are intended to explore patterns of change in the IAT that may have theoretical importance, but the results are not intended to implicate any specific source or process for these patterns of change.

The motivating questions behind this series of analyses are threefold: First, is performance in the IAT a function of time, such that IAT scores change across a session? Second, is IAT performance amenable to time-evolving analyses identifying the contribution of various stimulus attributes? Third,

does a dynamic approach to the IAT allow for a better understanding of links between it and other measures, such as survey measures or behavioral measures?

First, using a large openly available data source, the question of time-dependent systematic changes in IAT performance is explored. Next the relations between stimulus types and time are examined on a by-trial basis. Last, by-trial modeling of IAT performance is compared to aggregated measures in the context of predicting real-world behaviors.

Methods

A broad overview of IAT methods is provided here. More specificity is provided in each study's section. The IAT involves a speeded 2 alternative forced choice in which participants must rapidly choose to press a button associated with stimuli on a computer screen.

IAT were implemented in a standardized format with 6 blocks:

1. Single-stimulus training to associate response buttons with stimulus categories (e.g., female faces or words associated with competence)
2. Single-stimulus training to associate response buttons with stimulus categories
3. 20 trials of "practice" of dual-stimulus task pairing (e.g., Good+Black or Good+White)
4. 40 trials of dual-stimulus task pairing (e.g., Good+Black or Good+White)
5. Single-stimulus training to associate response buttons with stimulus categories
6. 20 trials of "practice" of dual-stimulus task pairing opposite of (3) and (4)
7. 40 trials of dual-stimulus task pairing opposite of (3) and (4)

For the current purposes blocks 3-4 and blocks 6-7 were combined into single 60-trial blocks over which change in IAT subtractive score may change. Study 1 reported here utilized open data from a large-scale and well-known online experiment (Xu et al., 2014). Study 2 is a re-analysis of by-trial data from Cox (2015; Experiment 2). Study 3 is a re-analysis of by-trial data, and many other measures, from Cox (2015; Experiment 3).

Study 4.3a: In aggregated IAT data, time-dependent changes are evident

Data source and description

To initially test whether variation in IAT scores may be time-dependent, Study 1 consisted of data downloaded from a publicly available data repository. Aggregated open data from 2005 was used (downloaded from https://osf.io/52qxl/). This is a large publicly available dataset from the well-known **Project Implicit** (Xu et al., 2014). The choice of one year was arbitrary but allowed a sufficiently limited data quantity to keep computational demands low. Experimental procedure followed the standard IAT outlined in the Methods.

The transparency and power inherent in a large publicly-available set of results provided a good starting point for testing time-dependent patterns of IAT scores. Unfortunately the publicly available data was only provided on an aggregated level, with several measures of performance extracted from blocks of 20 or 40 trials. While this aggregated data did not allow for a fully continuous-time approach to behavior within the IAT, the presence of time-dependent trends in aggregated measure would bely underlying continuous change and would motivate further examination of continuous change (see Study 3.2). The conventional approach to interpreting IAT effects is that an individual's response time and accuracy on each trial type is theoretically stationary across the duration of the experiment. This hypothesis regarding the generative process underlying IAT performance would be challenged by the observation of reliable time-dependent trends in response times that led to variations in calculated scores on the IAT. I test for these trends in this section while recognizing that the analysis is coarse due to the aggregated nature of the dataset. The following studies address the issues of aggregation.

Analysis description

Participants were first excluded for incomplete sessions or participants for sessions that were not the first time completing the IAT. This left 130799 participants. Next participants (n= 37627) were excluded for accuracy below 70% or mean response time above 1.5 seconds on any block of trials, leaving 93172 participants in the final analysis. While variation outside these exclusion criteria may be of interest in some contexts, the current goal was to include only participants who were very likely to be putting a good-faith effort into the task.

The aggregated IAT dataset had 6 variables of interest:

1. Session [participant] D [difference] score corresponding the canonical aggregated IAT effect
2. The order in which conditions were presented (i.e., Black+Good first or White+Good first)
3. Four response time means, for the 3rd, 4th, 6th, and 7th blocks.

Analyses leveraged two aspects of the experimental structure. First, the 20-trial practice blocks were identical to the succeeding 40-trial experimental blocks, allowing direct tests of practice effects on IAT response times. Second, across participants block types were counterbalanced, allowing comparisons of compatible and incompatible trials when participants had little experience or many trials of experience. Analyses here include (1) within-participant tests of change from practice to experimental blocks, and (2) between-participant tests of trial-type response time distributions separated by block number. That is, 3 comparisons addressed questions regarding the time-dependence of IAT results:

1. Were overall D systematically changed by the order of condition presentation?
2. Were RT systematically different across blocks of identical trials within participants? (i.e., 3 vs. 4 and 6 vs. 7)
3. Were effect sizes of condition differences systematically different between blocks 3, 4, 6, and 7?

Results

Overall IAT D scores were significantly related to the order in which the conditions are presented (see Figure 13; $T = -43.25$, $mean_{congruent\_first} = 0.294$, $mean_{congruent\_second} = 0.405$, $CI_{diff\_order} = [-0.116, -0.106]$, $d_{Cohen} = -0.28$). This indicated that time-dependent dynamics influenced participants' behaviors such that initial performance with White+Good stimuli led to smaller D scores than initial experience with Black+Good stimuli.

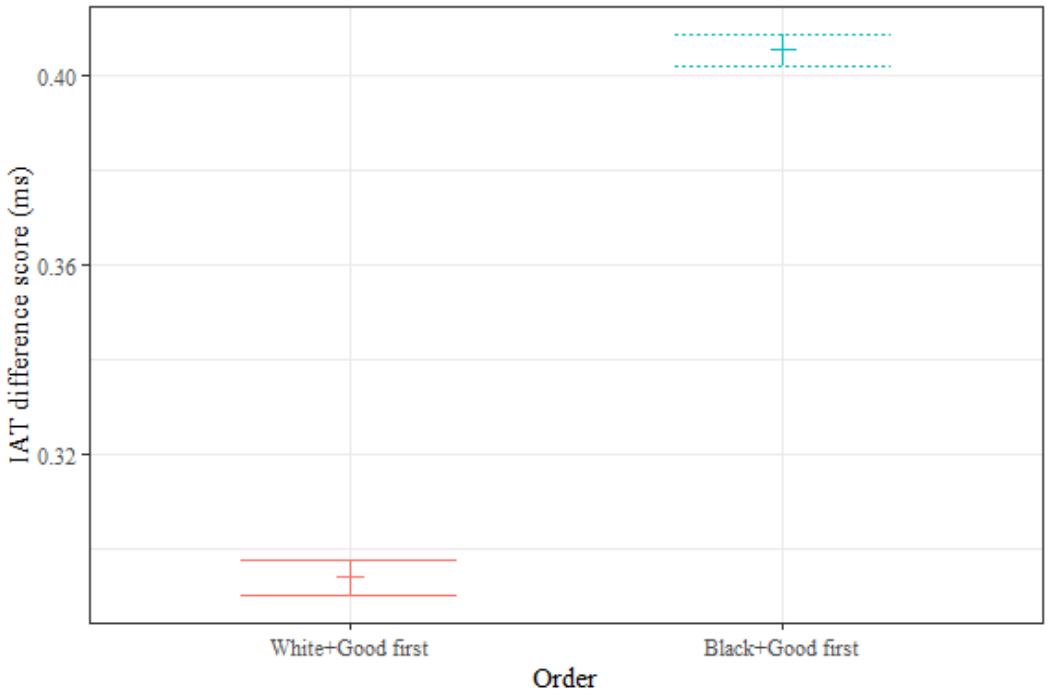


Figure 13. Standard IAT D scores indicating the magnitude of effects in the IAT, provided by ***Project Implicit****, separated by experimental order. The highly reliable difference indicates that the magnitudes of the IAT compatibility effect are highly influenced by the order of stimulus presentation. In turn, these order effects indicate within-task learning that differentially influences behavior in response to specific experiences that diverge between orders. Lines indicate means and bootstrapped 95% CI.*

Within participants, overall response times decrease from block 3 to block 4 (T = -304.83, $mean_{diff\_4\text{-}3}$ = -186, CI = [-187.2, -184.8], d_{Cohen} = -2) as well as from block 6 to block 7 (T = -197.54, $mean_{diff\_7\text{-}6}$ = -109.6, $CI_{diff\_7\text{-}6}$ = [-110.7, -108.6], d_{Cohen} = -1.29). Effect sizes were large in both cases. The changing pattern can also be seen in within-block between-subjects contrasts of response times (i.e., block-wise tests of the difference between RTs of White+Good First vs Black+Good First participants). The effects are large enough that error bars or T values would convey little information. Instead, in Figure 14 I show Cohen's d on each experimental block, providing a standardized estimate of the differences

between trial types by block. The pattern of change is nonmonotonic. Specifically, the most discriminative block was the first block of the second response pattern (i.e., highest effect size in block 6).

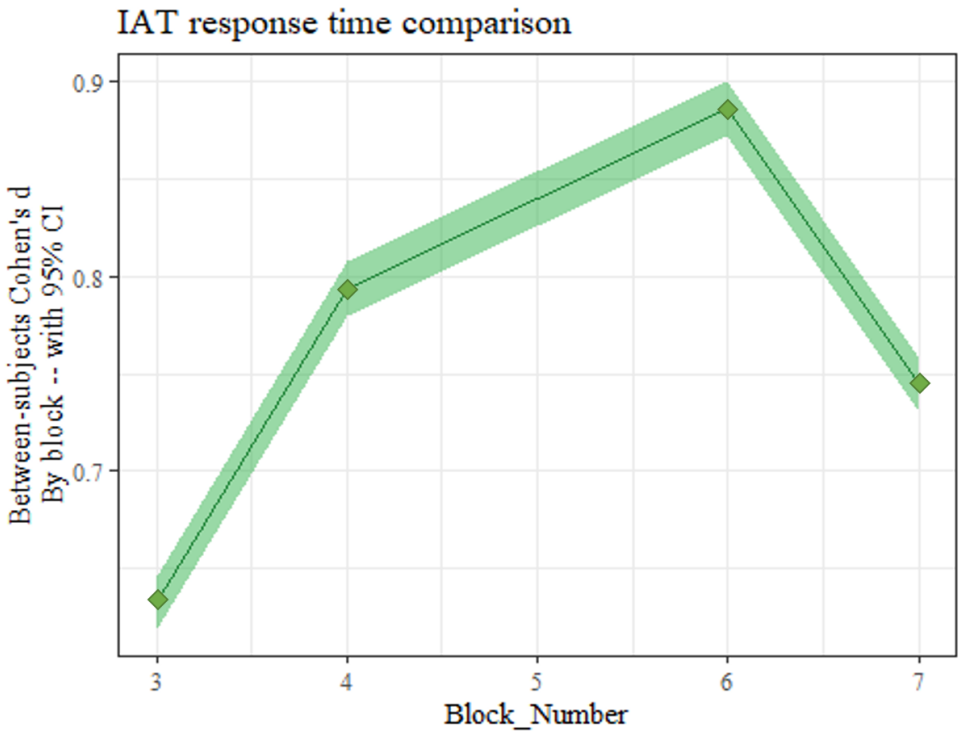


Figure 14. Between-subjects Cohen's d of IAT response time differences (i.e., standardized difference between distributions of compatible and incompatible response times in block 3, block 4, block 6, and block 7) The effect size of the between-subjects difference between compatible and incompatible stimuli changes as the task progresses, providing further evidence for time-dependent changes such as learning. Effect size increases from block 3 to 4 but decreases from block 6 to 7. Note that block 5 does not have dual-stimulus data and is not analyzed here. Shaded area indicates 95% CI of Cohen's d.

Discussion

Every result reported here indicates change in IAT results as a function of time within the task. More specifically, trials late in a session were systematically different from identical trials earlier in the same session. Differences were evident in both decreasing response time distributions and in the

asymmetric order-dependent sizes of IAT subtracted effects. Asymmetric IAT effects due to order implicate time-dependent processes (e.g., learning, self-regulation) that are not addressed by conventional aggregated approaches to the IAT. To overcome this limitation, in the following two studies I first demonstrate a method for fitting continuous IAT difference scores using by-trial hierarchical nonlinear regression. I next show that the relations between real-world measures of behavior and the IAT can be attributed largely to particular aspects of changing IAT difference values, thereby providing mechanistic specificity to observed effects.

Study 4.3b: In trial-wise IAT data, learning effects and order effects are evident

Data source and description

Data and methods have previously been reported in Cox (2015; Experiment 2). The reader is referred to that manuscript for full methodological details. This experiment involved 175 participants each completing a standard Black/White + Good/Bad IAT. In this and the following experiment, lower RT cutoffs were iteratively determined by testing the lowest RT on which participants performed above chance (Ratcliff & Tuerlinckx, 2002).

Hierarchical nonlinear modeling details

Response times tend to be well fit by skewed distributions, with the exponentially modified Gaussian distribution being particularly well-suited to analysis and interpretation of response times (e.g., Palmer et al., 2011; Ratcliff & Murdock, 1976). Data were therefore modeled as arising from an ex-Gaussian distribution. This distributional characterization of response time data includes three parameters: (1) the mean of a Gaussian component, (2) the variance of a Gaussian component, and (3) the scale parameter of an exponentially-distributed component, corresponding to the mean (and proportional to the variance) of that component. By using the exponential component of an ex-Gaussian distribution as the outcome of regression models (while controlling for other distributional parameters as constants, as with most regression approaches), predicted effects inherently incorporate the ubiquitous RT pattern of a covariance between means and variances. That is, predicting changes in the exponential component

allows for changes in a single parameter to lead to larger variances accompanying larger means, and vice versa.

Here response times were modeled as an ex-Gaussian distribution with a theoretical exponentially-distributed decision process overlayed by a Gaussian mean and variance of additive noise (see Appendix 7 for model formula). As response times were allowed to parametrically vary (e.g., due to stimulus type or as a function of time), these parameters modified the exponential component. Using ex-Gaussian nonlinear multilevel regression, RTs for compatible and incompatible trials were modeled as exponentially saturating functions of time. By utilizing both "training" and "non-training" blocks of IAT trials, this provided two runs of 60 trials for each participant's IAT data. The three parameters of the exponential learning function (starting level, rate of change, and asymptotic level) were each allowed to vary within each participant by (A) whether the target stimulus was either a white face or a positive-valence word, (B) whether the block's stimulus pairing was "Black+good/White+bad" or "White+good/Black+bad," and whether the block of trials was the second or the first. Nonlinear mixed-effects Bayesian models fitting utilized **rstan** via the **brms** package in R (Bürkner, 2017). Start, rate, and asymptote parameters were each sampled on log scales

Results

There was a modest decrease in the distribution of response differences (i.e., "Black+good or White+bad" vs. "White+good or Black+bad") from early trials to late trials (d_{Cohen} = -0.57). This can be seen in the difference in fit values of trial types (Figure 15) as well as in the difference in compatibility parameters of Asymptote vs `start` (Figure 16). Notably, there are no differences in rate of change across trial type or presentation order. In contrast, but in line with the aggregated analyses reported above, asymptotic response time is lower when Black+Good/White+bad condition is first, and this condition order is associated with a smaller asymptotic trial-type difference score. The latter effect is not reliably different than 0.

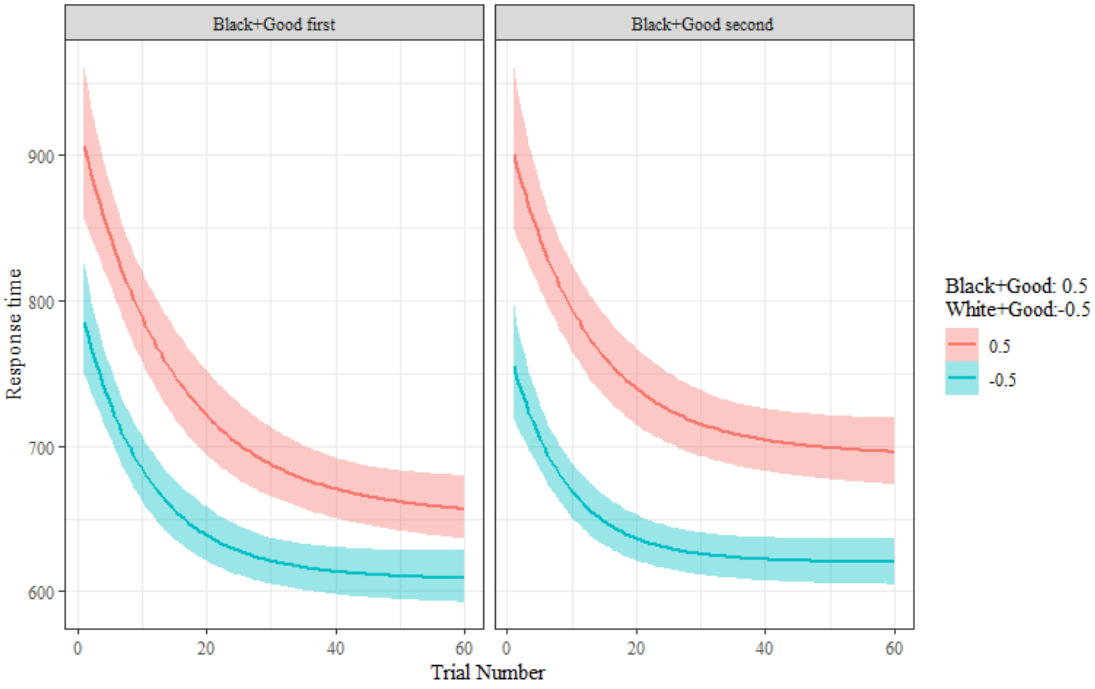


Figure 15. Fit values of response times on compatible and incompatible trials in Study 4.3b. Response times decrease over the course of each block, with a smaller decrease in incompatible trials when in the second block (i.e., participants in right panel have higher asymptotic RT than participants in left panel). Mean and 95% CI of fit values indicated.

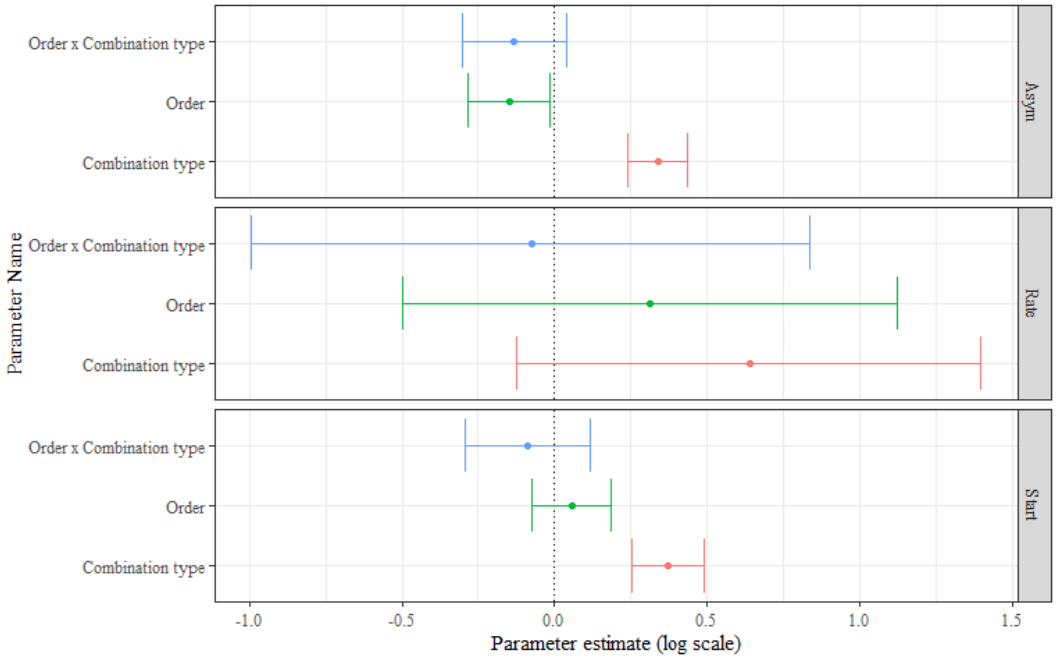


Figure 16. Fit parameters of Study 4.3b. Combination type indicates effect of incompatibility (i.e., "IAT effect"), for example the difference between White+Good and Black+Good. Order indicates whether a participant completed their incompatible or their compatible blocks first. Apart from main effects of incompatibility, only the main effect of Order on asymptotic RT was reliable. These results indicated that the largest and most time-related changes were standard learning effects (i.e., overall decreases in RT). While the interactions between Order and Combination type in predicting Asymptotic RT and Starting RT replicated the direction of Study 4.3.a, the effects were not reliable in this sample of 175 participants.

Discussion

As with aggregated results (Study 4.3a), by-trials models of the IAT demonstrated the systematic development of IAT differences scores within a standard task implementation. As response times decreased with task experience, the absolute magnitude of the IAT difference effect likewise decreased. This served as a proof-of-concept regarding model parameterization and implementation as well as baseline effects in time-related IAT change. In particular, trial type presentation order primarily influenced overall differences in RT asymptote.

One important aspect of the IAT is its potential to identify individual differences in implicit bias that may also be present in more ecologically valid contexts or behaviors. Variations between participants in IAT scores may arise from disparate sources, however, as indicated by Studies 4.3a-b. Individual-level variation in IAT performance, as modeled in Study 4.2b, could arise from variation in initial IAT score, rate of change in IAT score, or asymptotic IAT score. Each of the three components of changing IAT scores could be indicative of separate processes (e.g., rate of change may be related to self-regulation of prejudicial behaviors). In the following experiment I explore the degree to which variations in time-related IAT score parameters may be related to other measures of bias such as behavioral discomfort when discussing sensitive topics. None of these specific tests would be possible using standard aggregated approaches to individual differences in IAT performance.

Study 4.3c: Time-dependent measures of IAT provide specificity in correlations with other measures

Data and methods have been previously reported in Cox (2015; Experiment 3). The reader is referred to that manuscript for full details. All participants completed 4 IATs, 2 race-related IATs and 2 gender-related IATs. Participants also completed a set of surveys (e.g., external [EMS] and internal [IMS] motivation to respond without prejudice), and a structured in-person interview with an experimenter. The race and gender of the interviewer varied across participants. The interviews covered topics relating to race and gender. Behavioral variables were identified from the interviews, such as seating distance from the interviewer and third-party ratings of participants' seeming sexist or seeming racist (see Cox, 2015, for full description).

Cox (2015; Experiment 3) examined possible interactions between IAT scores, behavioral variables, and motivations to respond without prejudice. Results indicated that the relations between IAT performance and prejudicial behaviors were moderated by specific (i.e., either internal or external) motivations to seem less prejudiced. These links indicated that variations in participants' motivations to seem less prejudiced may determine the extent to which implicit bias is downregulated (leading to unbiased overt behavior) or not (allowing implicit and explicit measures to be correlated). Moderating

effects of motivations to seem less prejudiced thereby implicate a process of self-regulation that should be related to specific components of a continuous-time decomposition of IAT performance. Specifically, cross-participant variations in regulatory processes are most likely to relate to the rate of change in IAT effects (i.e., the rate of regulation). Regulatory processes may also be indirectly related to the asymptotic IAT effects (i.e., very effective regulation would result in lower IAT effects). In contrast, starting IAT effects should be unrelated to regulatory processes (i.e., IAT score prior to regulation).

Here I used a continuous-time model to decompose IAT performance and reconsider the primary results of Cox (2015; Experiment 3). The following patterns were reported using aggregated measures derived from the same dataset, and I ran the same analyses using time-dependent components of IAT performance.

- 3-way interaction between IAT, IMS, and experimenter race when predicting *seeming racist* (rated from behavior during interview) – due primarily to variation within the white-experimenter condition
- 3-way interaction between IAT, EMS, and experimenter race when predicting seating distance – due primarily to variation within the black-experimenter condition
- a 4-way interaction between IAT, IMS, EMS, and experimenter gender in predicting *seeming sexist* (rated from behavior during interview) – due primarily to a 3-way interaction within female-experimenter condition

Cox (2015; Experiment 3) tested each of these models in the full sample, including interactions with experimenter demographics, and then re-tested in the relevant subsample. I limited my scope to these relations. IAT response times were modeled, identically to Study 2, using hierarchical nonlinear ex-Gaussian regressions (see Appendix 7 for model formula). Race and Sex IATs were each fit with single models, with IAT type (e.g., competent, good, leader) coded as -0.5 and 0.5 for the purposes of estimating overall parameters at levels intermediate between the IAT types. Participant-level parameters were then

extracted from the mixed-effects model, and participant-level parameters corresponding to differences between trial types were used in subsequent analyses predicting non-IAT measures.

Robust linear models using the R package **MASS** were used to test relations between measures. Exact degrees of freedom cannot be not clear given the iteratively reweighted least squares method of robust linear models, but even at a very low breakdown point of 50% of observations (df = 72) the T value associated with a $p < .05$ would be approximately 1.99. As such, T values of 2 were heuristically used as thresholds for reliability.

Current analyses

In order to test individual differences, from the initial 188 participants I excluded people for several reasons. 28 participants were excluded for missing variables of interest (e.g., EMS scores). An additional 19 were excluded for being multivariate outliers among variables of interest. This was determined using a robust Mahalanobis distance method (Leys et al., 2018). Robust covariance estimation used a minimum 90% of cases, and outlier rejection used an alpha of .01. Participant rejection left 141 participants remaining.

Parameters and descriptions and results

Fit RT values of the generalized nonlinear mixed-effects models are shown in Figure 17.

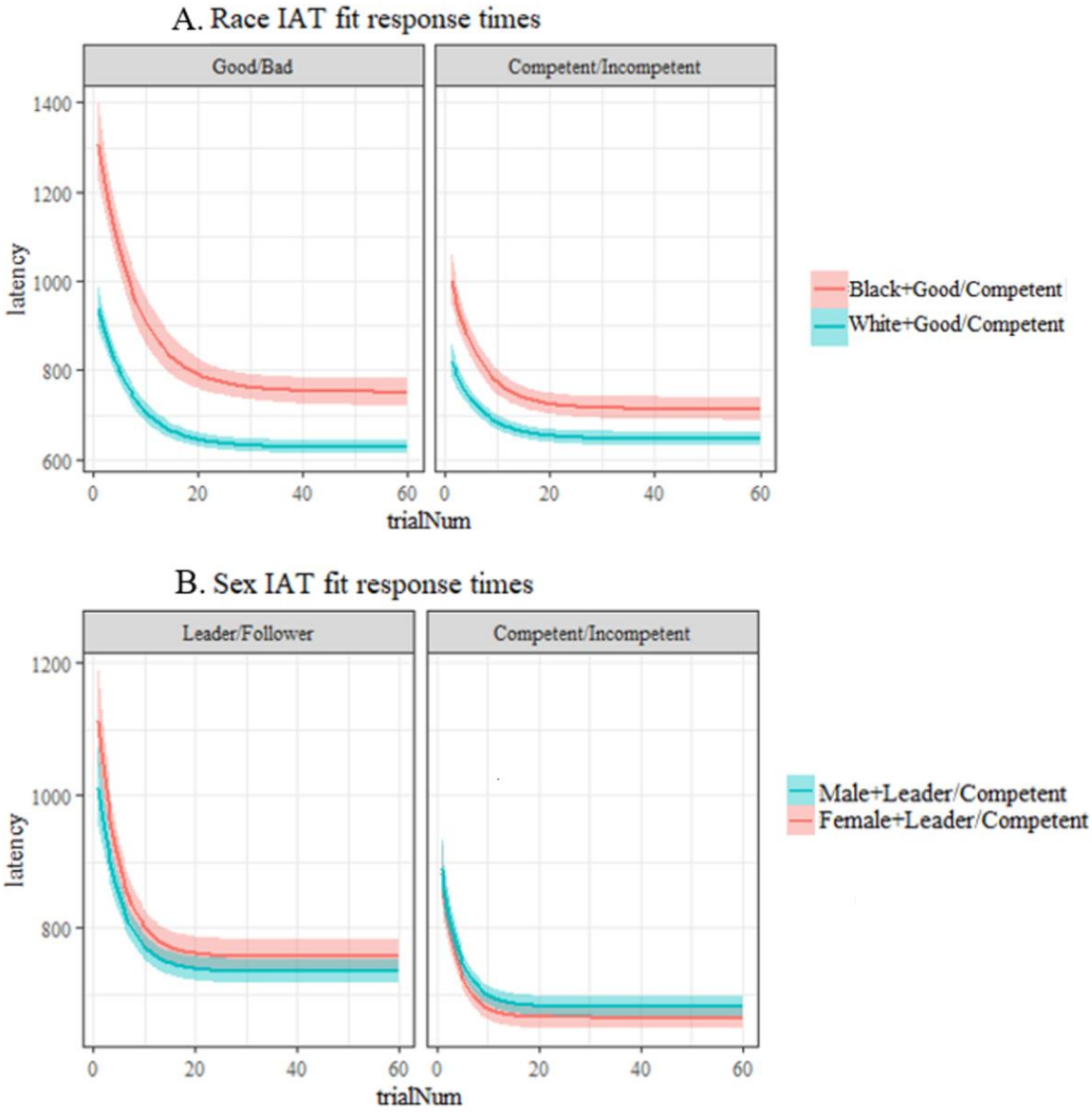


Figure 17. Plots of Study 4.3c model fits. (A) Race IAT trial type difference scores (i.e., compatibility effects) were reliable, while (B) Sex IAT trial type difference scores were not reliable. The latter pattern was due to a reversal in response time patterns across the two types of Sex IAT; while `Female+Leader` *was slower than* `Male+Leader`*,* `Female+Competent` *was faster than* `Male+Competent`*. Due to the ambivalent Sex IAT pattern and the overall interest in bias, only subject-level effects for the* `Leader/Follower` *IAT were used in subsequent analyses. Note that this use of a single IAT parallels*

the analyses of Cox (2015, Experiment 3). In contrast, overall subject-level Race IAT trial type effects are used in subsequent analyses because the effect of trial type was evident, and the same numerical direction, in both Race IAT types. Still, it is notable that the `Good/Bad` *Race IAT showed larger trial type effects than the* `Competent/Incompetent` *Race IAT.*

Analyses following Cox (2015; Experiment 3)

Analyses, testing the findings described above from Cox (2015; Experiment 3), were separately conducted using 4 IAT-derived measures (i.e., overall score and 3 time-dependent parameters). That is, the IAT scores entered into the original models were either overall measures or single parameters describing aspects of the time-evolving IAT score. Each effect associated with an IAT score was compared to a critical T value of 2.

All plots below organize different models' parameters into figure panels, with the measure of IAT (and likewise the specific model) color-coded. Plots show T values of model coefficients, with reference lines drawn at +2 and -2 to indicate thresholds of reliability. No reliable effects were observed in predicting Seeming Racist (see Figure 18 and Figure 19). In contrast, rate of change in the magnitude of RT differences interacted with EMS to predict seating distance from the experimenter (see Figure 20 and Figure 21).

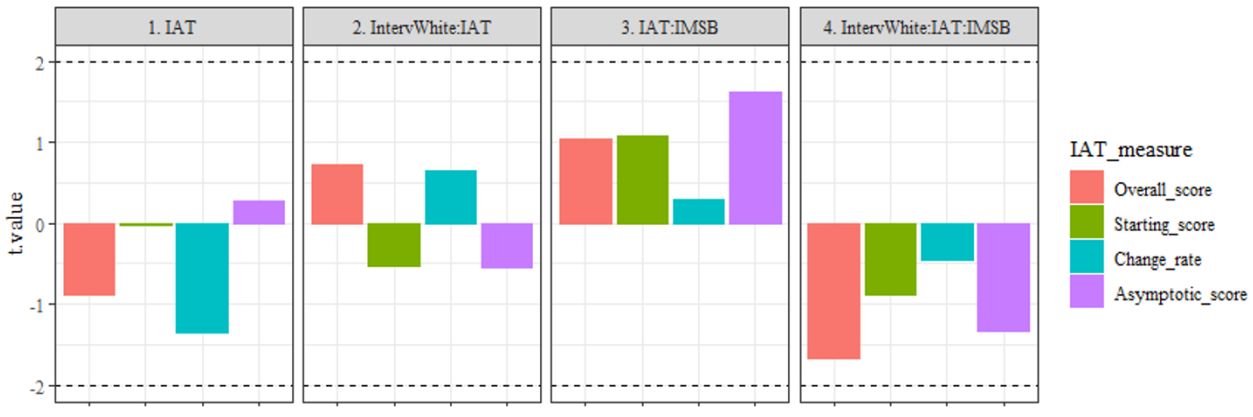


Figure 18. T values of IAT measures' predictiveness of seeming racist. Results of 4 models, organized by coefficients including IAT. No IAT measure had any reliable main effect or interaction. Dashed lines indicates |T|=2, the threshold for reliability used here.

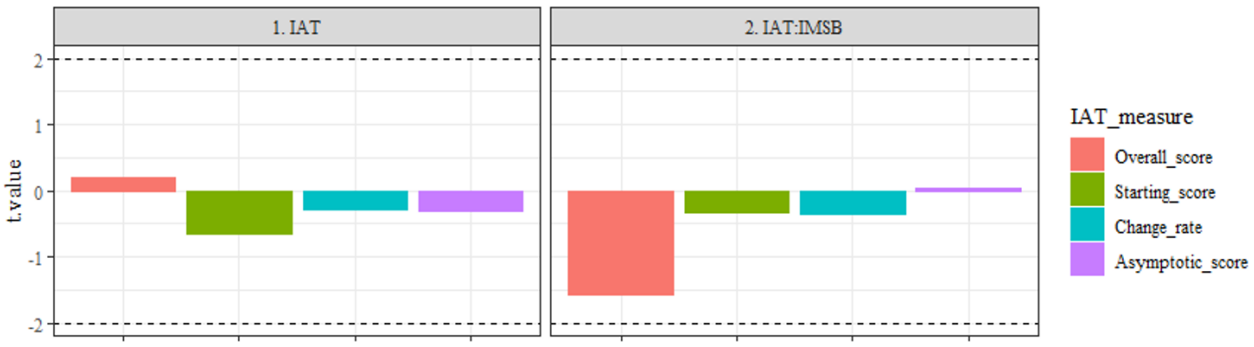


Figure 19. T values of IAT measures' predictiveness of seeming racist within participants whose experimenter was White. T values of 4 models, organized by coefficients including IAT. No IAT measure had any reliable main effect or interaction. Dashed lines indicates |T|=2, the threshold for reliability used here.

Figure 20. T values of IAT measures' predictiveness of seating distance within all participants. T values of 4 models are shown, organized by coefficients including IAT. Overall IAT compatibility differences interacted with EMS scores, as well as interviewer race, in predicting seat distance from the interviewer. This effect was likewise reliable when using rate of change as the index of IAT performance but not when

using starting or asymptotic RT difference as indices of IAT performance. Dashed lines indicates |T|=2, the threshold for reliability used here.

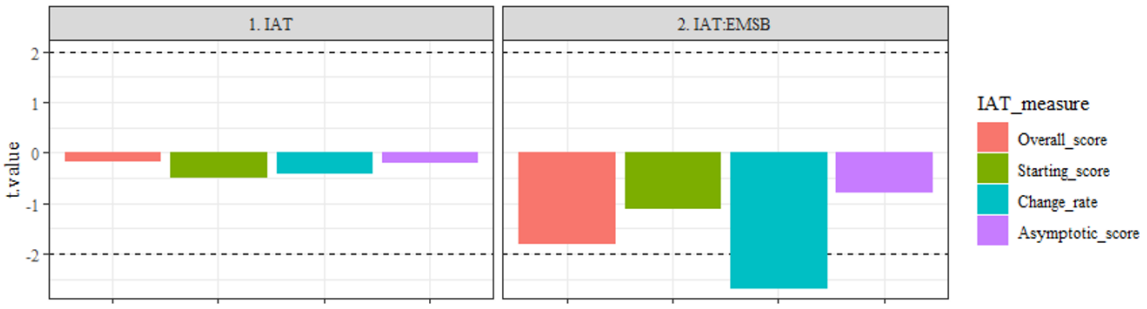


Figure 21. T values of IAT measures' predictiveness of seating distance within participants whose experimenter was Black. T values of 4 models are shown, organized by coefficients including IAT. Only the rate of change in IAT effect, interacting with EMS, showed any reliable predictiveness of seating distance. Dashed lines indicate |T|=2, the threshold for reliability used here.

The results of the Race IAT from Cox (2015, Experiment 3) were partially replicated, with the IAT rate of change paralleling to the observed overall effects. Next I utilized the same methods to test the various components of the Sex IAT results of Cox (2015, Experiment 3). Models predicting Seeming Sexist were only reliable within participants with Female experimenters (see Figures 22 and 23).

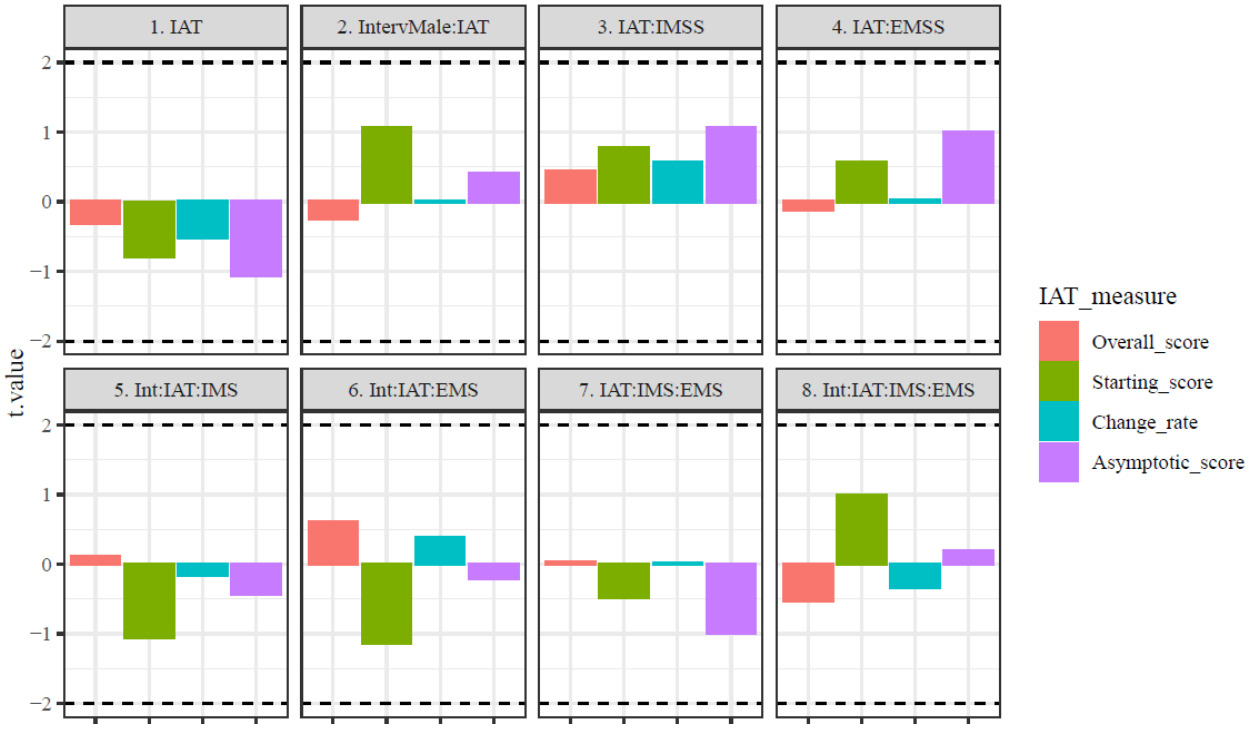


Figure 22. T values of IAT measures' predictiveness of seeming sexist among all participants. T values of 4 models, organized by coefficients including IAT. No index of IAT effect was reliably related to Seeming Sexist. Dashed lines indicate |T|=2, the threshold for reliability used here.

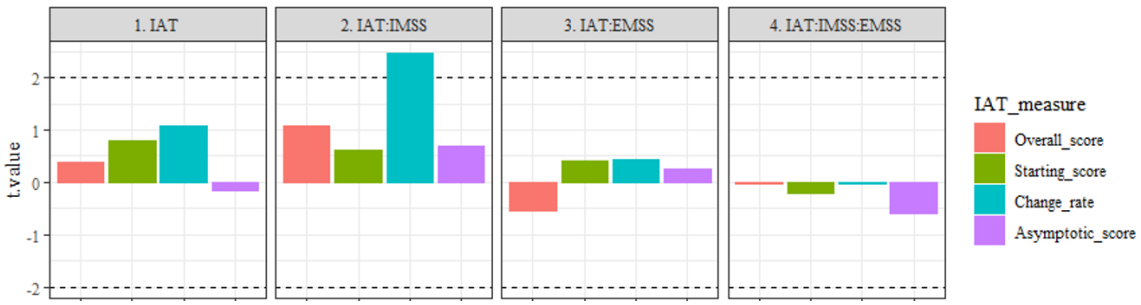


Figure 23. T values of IAT measures' predictiveness of seeming sexist within participants whose experimenter was Female. T values of 4 models, organized by coefficients including IAT. When only analyzing only data from participants with a female interviewer the interaction between IMS and rate of IAT effect change was a reliable predictor of Seeming Sexist. Dashed lines indicate |T|=2, the threshold for reliability used here.

Study 4.3c results summary

The most striking result is that Race IAT score *rates of change* interacted with external motivation to seem less prejudiced in predicting a behavioral measure (seat distance). Sex IAT score *rates of change* interacted with internal motivation to seem less prejudiced to predict Seeming Sexist in contexts with female interviewers. In aggregate these results indicated the potential utility of using time-evolving measures to clarify the relations between IAT scores and real-world behaviors.

Study 4.3 discussion

Across 3 experiments IAT effects demonstrated robust within-person changes. Learning, in the form of decreased response times, was evident in an aggregated large dataset as well as using by-trial modeling methods. The uniformity of these learning effects should prompt an investigation of the possibly independent processes by which certain aspects of IAT effects, such as initially large incompatibility effects or rapid suppression thereof, might be differentially informative to theories of implicit bias.

When correlating components of participants' learning curves with behavioral measures, there was the most evidence for the importance rate of change in compatibility effect. This implicates possible individual differences in the efficacy of self-regulation processes, as opposed to initial "gut reactions" or asymptotic "persistent bias." For instance, in replicating the analyses of Cox (2015; Experiment 3), rate of change was qualitatively similar to overall IAT effect in its interaction with External Motivation to Seem less prejudiced (EMS) to predict seating distance from an experimenter, particularly when the experimenter was black. Prior interactions between IAT and IMS/EMS, in predicting real-world behaviors, have been interpreted as evidence for self-regulation processes that may either allow or prevent implicit bias from becoming explicit bias. The results I reported here support this interpretation due to the uniformity of reliable effects of a regulation-like component of IAT performance (i.e., rate of change). This mechanistic specificity was impossible in the absence of a time-evolving model of the IAT.

In Chapter 4 the demonstrations of learning in both IAT and Flanker tasks have shown that, despite the conventional implicit assumption of stationarity in the processes giving rise to behavior each paradigm, systematic changes occur as participants learn within the task contexts. Few implications of this timescale of change were explored (but see Study 4.3c). Instead, this chapter has provided initial evidence regarding the potential benefits of a learning-centered approach to experimental psychology. Chapter 5 contains further examples, including an examination of individual differences in learning over the course of just 64 trials.

Chapter 5. A learning-centered approach to individual differences

Previous chapters have demonstrated benefits of time-continuous models of learning tasks (Chapter 3) as well as the applications of a learning-centered approach to response competition tasks in the contexts of both experimental social and cognitive psychology (Chapter 4). In this chapter I turn to applications of a learning perspective to correlational studies of psychological functions. Patterns of correlations within psychological data can provide great insight into the interdependence of processes and the broader structure of mental abilities (Colquitt et al., 2000; Kane et al., 2004; Miyake, 2000; Pearl, 2009). Intra-individual variation has frequently been treated as noise, however (Molenaar, 2004), a fact which is belied by the very label "individual differences." To the extent that intra-individual variation has been considered (e.g., across timescales of weeks, months or years), short-term trajectories of change remain to be considered a nuisance (e.g., a sign of a measure's unreliability). In contrast, in this chapter I show the utility of learning estimates that capture individual differences in by-trial monotonic changes using two paradigms. The first concerns personality, lifestyle, and other predictors of learning on two perceptual learning tasks. The second addresses the mechanistic bases of correlations between fluid intelligence measures and working memory capacity.

Study 5.1 Individual differences in perceptual learning

With dedicated training most individuals can improve abilities as complex as playing chess or as simple as making basic perceptual judgements. Nonetheless, in response to identical training regimes, there is a wide range of possible outcomes. Some learners may improve substantially while others show few or no benefits in response to training. Although inter-individual variation of learning has been explored in various contexts (Ackerman & Cianciolo, 2000; Burke & Hutchins, 2007; Jaeggi et al., 2014), previous studies have largely utilized aggregated estimates of performance that limited the possible inferences (see Chapter 2 & Kattner, Cochrane, Cox, et al., 2017). Consideration of individual-level rates of learning, as dissociated with other parts of learning trajectories, provides the ability to empirically

distinguish between theoretically distinct mechanisms of generalization (i.e., immediate benefit to a novel task or improvement in the ability to learn a novel task).

My colleagues and I first established that, by using models of learning that were sensitive to by-trial changes in performance, we could identify reliable inter-individual variations in components of learning. Perceptual learning is typically studied using methods highly specific to specific experimental paradigm. Tests of generalization tend to use stimuli with different properties within the same experimental paradigm (e.g., Study 3.3). Correlations between perceptual learning performance is not a conventional topic of study, though, and tests of generalization are instead of interest when multiple learning tasks are studied in conjunction (e.g., Study 3.2). However, there are various reasons to believe that there should be between-person differences in abilities that would lead to systematic variations in perceptual learning. Individual differences in learning ability are predicted from broad theories of intelligence and learning (Ackerman & Cianciolo, 2000) as well as specific accounts of learning-to-learn through the ability to adapt perceptual templates (Bejjanki et al., 2014), yet the nature of these individual differences has not been addressed in the field of perceptual learning. Additionally, continuous-time models of learning allow for greater specificity in characterizing the patterns of relations between experimental paradigms (e.g., a lack of initial correlations would demonstrate a lack of preexisting abilities, and the presence of rate-of-learning correlations would demonstrate a common ability-to-learn mechanism). These patterns are indeed what we observed in Study 5.1.

Participants completed two perceptual learning tasks over the course of several days each. We found that certain components of their learning curves were reliably correlated across tasks. We next established that, above and beyond the reliable individual differences during initial training, learners varied in their abilities to generalize learning to novel stimuli. After establishing reliable inter-individual variations in learning as well as generalization we tested for the relations between various these learning components and measures that may explain inter-individual variations in learning (e.g., speed of processing, working memory, conscientiousness). While the patterns of relations were quite heterogeneous, we did observe confirmation of certain predicted relations (e.g., initial performance was

reliably related to mean response times on a simple speeded task) as well as disconfirming other predictions (e.g., reasoning scores were not related to aspects of learning; see Appendix 8).

Study 5.2 Individual differences in fluid intelligence and working memory

In Study 5.1 my colleagues and I demonstrated individual differences between specific components of learning (e.g., rate) and various possible predictors of learning (e.g., dispositional or cognitive abilities). These results are evidence for the utility of approaching learning from a componential time-dependent perspective in order to identify the bases of cross-person variation. Study 5.1 only approached the putative "learning" tasks in a time-continuous manner, however, even though other aggregated measurements may have involved learning as well (e.g., response time tasks, working memory task). As shown in Study 4.3c, systematic and theoretically meaningful patterns of learning may arise over the course of a few dozen trials. Although we aggregated many measures in Study 5.1 in order to limit the number of predictors tested, we thereby accepted a conflation of various possible sources of between-person variation (see Chapter 6 for additional discussion). Individual differences need not take this step of aggregating measures such as working memory. In Study 5.2 I use a continuous-time approach to modeling working memory performance and correlations with fluid intelligence.

Fluid intelligence has been, from its origins as a construct, defined as a latent ability related to successful performance independent of prior experience (Horn & Cattell, 1966). It should be no surprise then that performance on a great number of tasks in cognitive psychology, which have largely been designed to measure processes independent of explicit knowledge or strategies, would be correlated with measures of fluid intelligence. A specific pattern of individual differences has gained a large amount of attention though due to consistently large covariations. Specifically, fluid intelligence and working memory capacity have been so closely linked that some researchers have discussed an isomorphism between the constructs (Kyllonen & Christal, 1990) or a developmental cascade linking them (Fry & Hale, 1996). A plethora of research into these connections followed the earliest demonstrations of these patterns, including multiple large studies using structural equation modeling to examine the putative process-level bases of links between memory measures and fluid intelligence measures (reviewed in Ackerman et al., 2005). Miyake et al. (2000) used structural equation models to establish the extremely influential tripartite division of "executive functions" into *shifting*, *updating* and *inhibition*, one or more

of which may be present in any given working memory task. Similarly, multivariate models have been used as evidence to relate intelligence specifically to the "capacity limitations" of working memory (Fukuda et al., 2010; Shipstead et al., 2014; Voelke et al., 2014).

There is serious concern that a broad covariance-based approach may obscure aspects of the mechanistic reasons for covariation. A different strain of research on the relationship between the two constructs has examined the demands of various tasks and items therein (Harrison et al., 2015; Jaeggi et al., 2010). This lower-level approach has provided complementary evidence to the large-scale latent-variable approach. Both have concluded that specific aspects of working memory, such as the brief storage and manipulation of memoranda, are likewise necessary aspects of successful performance of fluid intelligence tasks. While there are as many variations on this narrative as there are theories of working memory, the methods implemented by most authors have constrained mechanistic explanations to a common set of possibilities. However, the putative learning component that is at the very core of fluid intelligence is conspicuously absent from the body of work. The foundational conception of fluid intelligence as a predictor of success in novel areas parallels the intended design of cognitive tasks as measurements of experience-independent abilities. Yet the very ability of fluid intelligence to predict success in novel complex skills implies that the construct is also related to the ability to learn these new complex skills. That is, complex cognitive tasks must be learned in order to be completed correctly, and correlations with fluid intelligence may be due to features of the tasks themselves or due to the ability to learn those tasks.

The necessity of learning within cognitive tasks means that correlations between fluid intelligence and overall working memory performance (i.e., averaged across time) may be conflating three possibilities. The standard assumption may hold, and correlations may indicate shared stable process-level constraints or abilities. Alternatively, correlations may not indicate shared *stable* processes but instead indicate that higher fluid intelligence gives certain people immediate boosts to performance in difficult novel tasks, with lower fluid intelligence producing the opposite result. In support of this view, Ackerman

and Cianciolo (2000) demonstrated initially high correlations between fluid intelligence and performance in a complex multitasking environment, with correlations monotonically decreasing with task practice.

A last possibility is that variation in fluid intelligence would be related to learning itself. Fluid intelligence in relation to variation in learning ability, as a construct, has received comparatively little attention *per se* in studies of cognition. In contrast, the links between intelligence and learning ability within educational, occupational, and other "real-world" settings are widely reported and historically important (Binet & Simon, 1916; Cattell, 1963; Nesbitt et al., 2015; Primi et al., 2010; Ren et al., 2015; Vaci et al., 2019). Notably many positive relations have been reported between fluid intelligence measures and learning, with learning being characterized as the magnitude of change in performance. Positive relations in this context are indicative of a "Matthew effect" (Stanovich, 1986) in which individuals who have pre-existing advantages also acquire the most benefits. Above and beyond the associations between fluid intelligence and the magnitude of learning, and despite the wealth of justification for a theoretical distinction between the rates and magnitudes of learning (i.e., quantity learned vs. efficiency of learning; Bavelier et al., 2012; Bejjanki et al., 2014; Harlow, 1949; Kattner, Cochrane, Cox, et al., 2017), there have not been investigations of fluid intelligence as a predictor of variance in the rate of learning independently of learning magnitude.

Consistent theoretical and empirical links between learning and fluid intelligence have been reported, yet investigations of the cognitive basis of intelligence have largely avoided investigations of learning as a mechanistic link between intelligence measures and cognitive measures. While some results have shown that learning within fluid intelligence measures is an important component in its predictiveness of disparate measures (T. Wang et al., 2017), cognitive measures have not been treated as learning environments in which fluid intelligence may be an important source of inter-individual variation. Cognitive tasks are novel complex environments in which humans must learn, just like many other real-world contexts. I used this learning-centered perspective to test the extent to which a working memory task may be related to fluid intelligence due to various sources. First, variation in fluid intelligence may be directly related to the ability to perform novel tasks well. This aligns most closely to

the concept of fluid intelligence as “the ability to succeed on novel tasks” and would manifest as correlations between *initial* working memory accuracy and reasoning scores. Second, fluid intelligence may be related to the ability to learn novel tasks rapidly, which would align most closely with a construct of intelligence as “the ability to learn” and would manifest as correlations between reasoning scores and the *rate of change* in working memory accuracy. Third, fluid intelligence may be related to relatively stable processes necessary for the performance of both working memory tasks and reasoning tasks; this would align most closely with a definition of fluid intelligence as “sharing processing constraints with working memory” and would manifest as correlations between reasoning scores and *final* working memory accuracy. Finally, the *magnitude of learning* can be calculated as a derivative measure from the initial and final levels of performance. Each of these possibilities provides information that any measure averaging over all working memory trials could not provide.

It is important to note that any combination, or none, of the above links could be present. The possible links between working memory task performance and fluid intelligence task performance are not mutually exclusive. They are likewise not necessarily simply explaining a part of the relations between overall performance on the two tasks. To the extent that the possible mechanistic independence of the four possibilities is borne out in statistically independent parameter estimates it is also possible that there is a range in the extent to which links to fluid intelligence explain the overall links, vs demonstrate novel relations.

Here I tested the four candidate explanations for links between working memory and fluid intelligence. I decomposed working memory performance into by-trial estimates of accuracy changing as a continuous nonlinear function of time, thereby identifying components related to each of the four possibilities outlined above: Initial performance, final performance, rate of learning, and magnitude of learning.

Methods

Individuals were assigned to groups differing on the presence or absence of feedback (feedback-present: 55; feedback-absent: 60).

Procedure

All spatial span and matrix reasoning stimuli were presented in an Internet browser using the Qualtrics survey software. Participants used Google Chrome on Dell 22-inch monitors. Tasks were designed to last approximately 45 minutes total. Two types of tasks were completed by all participants. The first, matrix reasoning, consisted of two pattern-matching tasks modeled after Raven's Progressive Matrices (Raven, 1998). One set of matrices were a set of 14 items taken from the stimulus set of Matzen et al (2010) while the other set of matrices were a subset of 18 items from Pahor et al. (2019). Every trial of each matrix type was a pattern completion problem in which 8 items were presented in a 3x3 grid, with the bottom-right item missing. The 8 items contain some pattern(s) or rule(s) dictating the size, shape, number, pattern, color, or orientation of the elements of the missing item. Eight possible items to complete the pattern were provided for each trial.

The second task was spatial span modeled after a Corsi block-tapping task (Berch et al., 1998; Cochrane et al., 2020). On each trial 16 possible targets were outlined in black at random locations in an invisible 6 x 6 grid. After 500 ms, one square would fill with color for 1 second before returning to an empty outline. Immediately a different square would fill with color, and so on until the trial's set size was reached. A sequence of random non-repeating target locations was defined through these color changes. After the sequence participants used their mouse to reproduce the sequence of targets. In the feedback-present condition participants were provided with the percent correct they received on each trial; in the feedback-absent condition they were not given any information about their performance.

Set sizes were blocked in groups of four in the following order: four of set size 6, four of set size 8, four of set size 5, and four of set size 8. This series of 16 trials was repeated four times for a total of 64 trials. There were two very easy catch trials of set size 3 in between the first and second blocks and between the third and fourth blocks. There was a break between the second and third block (i.e., halfway through the task). Participants were not informed of the set size on any trial. Identical trial orders of set sizes ensured that participants' experience in the task was comparable and variations in learning would not be due to selection of trial difficulties.

Analysis

Participants' percent correct was calculated for each matrix reasoning task. Data was first screened for below-chance performance on either matrix reasoning task. Chance performance was defined as failing to be significantly above the guessing rate of 12.5%, utilizing a one-sided binomial test. This translated to excluding participants who scored 3 or lower on the Sandia matrices (Matzen et al., 2010) or 4 or lower on the UCMRT (Pahor et al., 2019). Matrix reasoning percent correct was Z-scored within each task, then these normalized results were averaged to produce the composite matrix reasoning score used in all results. Note that due to the idiosyncrasies of the matrix reasoning items no time-evolving analysis was attempted. That is, the trials were not *iid* or parametrically-varied samples of performance, thereby complicating the process of fitting a continuous function of change. While the fitting of this type of data may be the subject of a future study, it was considered beyond the scope of the current study. Chance performance was also assessed on spatial span catch trials of set size 3, however, no participants were excluded by this criterion.

Next, participants' spatial span data was fit with a generalized nonlinear Bayesian mixed-effects model using the R package **brms** (Bürkner, 2017; see Appendix 9 for model specification). Exponential change in accuracy was modeled in terms of individual-level [random effects] initial level of performance, asymptotic level of performance, and of amount of time taken to 50% of change (i.e., rate). Starting and asymptotic accuracy parameters were estimated on inverse-logit scales in order to provide unbounded parameter ranges associated with bounded ranges of predicted accuracies. Rate parameters were estimated on a binary log scale.

The fixed effect (i.e., between-subject) of feedback was fit for each parameter, and random slopes were fit for each participants' start and asymptote parameters (noting that random slopes were estimated within the inverse-logit transformation described above). Set size was centered at 6.5 to estimate effects at an intermediate difficulty. While participant-level estimates of rate, asymptote and start were the primary outcomes of interest (i.e., as potentially correlated to matrix reasoning), fixed effects of feedback on each nonlinear parameter were of secondary interest.

As initial tests of the bivariate relations between components of working memory performance and matrix reasoning accuracy a robust measure of correlation was used. Each of the two variables entered into the correlation calculation was first transformed by first finding the optimal Yeo-Johnson power transformation (i.e., lambda to minimize skew), applying that transformation, then calculating the Pearson product-moment correlation of 1,000 bootstrapped datasets (i.e., resampled with replacement). The median, 2.5% and 97.5% quantile are reported as the point estimate and confidence intervals for the relation. The T value of the median correlation was also converted to a log-2 Bayes factor using default values in the R package **BayesFactor** (Morey et al., 2015; Rouder et al., 2009). Log-2 Bayes factors have an intuitive interpretation regarding the evidence for the correlation or lack thereof; a reported BF_{log2} of 1 indicates twice as much evidence for the correlation as for a lack of correlation, while a BF_{log2} of -2 indicates four times as much evidence for the lack of correlation as for the presence of the correlation, and so on. Note that conventional interpretations often use a cutoff of 3 to indicate "more than anecdotal" evidence , which correspond to BF_{log2} smaller than -1.58 or larger than +1.58 (Wetzels et al., 2011). If Bayes factors from correlations indicated evidence for bivariate relations between a working memory component and matrix reasoning scores, then that component was used in subsequent regression models. All regressions were robust linear models fit using **TEfits** function **tef_rlm_boot**. Models use robust M-estimation, include bootstrapped within-sample robust confidence intervals, and report the median out-of-sample predictive change in error (i.e., ΔR^2_{oos}). Median out-of-sample predictive change in error was calculated through repeated resampling of 80% of data, fitting the robust linear model to that sample, using the model to predict the out-of-sample data, and testing the proportional reduction of error for the out-of-sample data.

Results

The feedback-present group initially included 55 participants; 12 were excluded for chance performance on the UCMRT, 6 were excluded for chance performance on Sandia matrices, and 0 were excluded for chance performance on set-size-3 spatial span. The feedback-absent group initially included 60 participants; 11 were excluded for chance performance on the UCMRT, 1 were excluded for chance

performance on the Sandia matrices, and 0 were excluded for chance performance on set-size-3 spatial span. This left 42 participants in the feedback-present group and 45 participants in the feedback-absent group.

All percent-accuracy fits were estimated on log-odds [logit] scale, and all analyses retain this transformation in order to satisfy the linearity implied in regressions and correlations. In brief checks for robustness it was confirmed that results were qualitatively similar when using raw proportions. Bootstrapped robust linear models used 5000 resamples for both bootstrapped CI and out-of-sample delta R-squared.

Changes in spatial span scores

Overall, there was a trend toward improvement from starting values to asymptotic values (m_{diff} = 0.17, T = 1.91, CI_{diff} = [-0.007, 0.347], d_{Cohen} = 0.412; see Figure 24). In total, 50 of the 87 participants' scores increased.

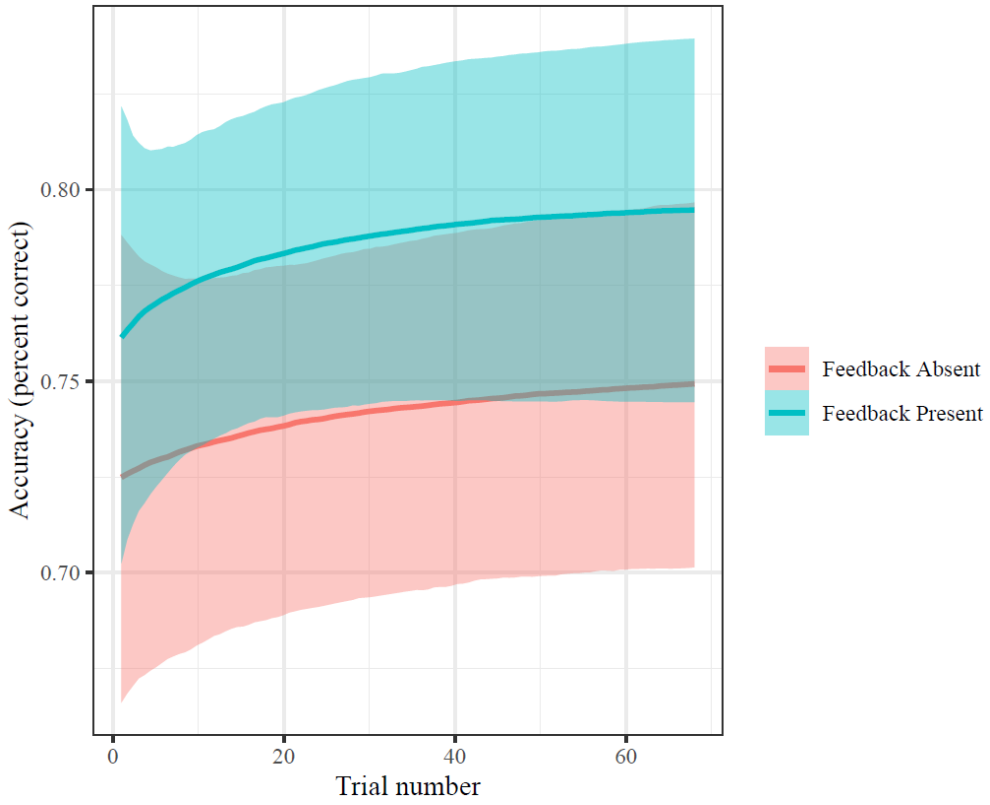


Figure 24. By-trial accuracy fit values of spatial span, fit at the mean set size (i.e., 6.5). Mean fit accuracies and model 95% CI are shown. Feedback resulted in a consistent benefit of approximately 3.5%. Although both feedback conditions showed trends toward increasing accuracy there was a large amount of variability.

I first tested differences in fit values by feedback condition. The presence of feedback was reliably associated with faster change in performance (b = -1.1, CI = [-1.8,-0.29], dR^2_{oos} = 0.0632) but not starting or asymptotic performance (start: b = 0.21, CI = [-0.077,0.48], dR^2_{oos} = 0.0209; asymptote: b = 0.21, CI = [-0.02,0.46], dR^2_{oos} = 0.0341).

Correlations in all data

There was a high correlation between the two matrix reasoning tasks (r(85) = 0.7 [0.57,0.8], BF_{log2} = 35.61), thereby justifying their inclusion into a single score by independently z-scoring them and then averaging these normalized values.

Overall accuracy was almost perfectly correlated with the estimated ending accuracy (r(85) = 0.96 [0.93,0.97], BF_{log2} = 141.37), highly correlated with starting accuracy (r(85) = 0.59 [0.43,0.73], BF_{log2} = 21.9), and less reliably related to rate of change (r(85) = -0.24 [-0.44,-0.02], BF_{log2} = 0.41; see Figure 25 for visualizations of simplified analyses utilizing Spearman correlations). The near-isomorphism between ending and overall accuracy provided a clear grounding for the interpretation of overall effects as overwhelmingly reflecting stable asymptotic aspects of the ability to correctly remember items, while the smaller correlation between ending and starting accuracy (r(85) = 0.38 [0.18,0.57], BF_{log2} = 6.26) than between either of these estimates and overall accuracy provided some assurance that the estimates were dissociating variation in performance that would typically be confounded within overall scores. The small-to-moderate magnitudes of associations between parameter estimates boded well for their interpretation, as they did not imply that the three measures were simply reflecting the same underlying variation despite being extracted from the same data as the overall accuracy.

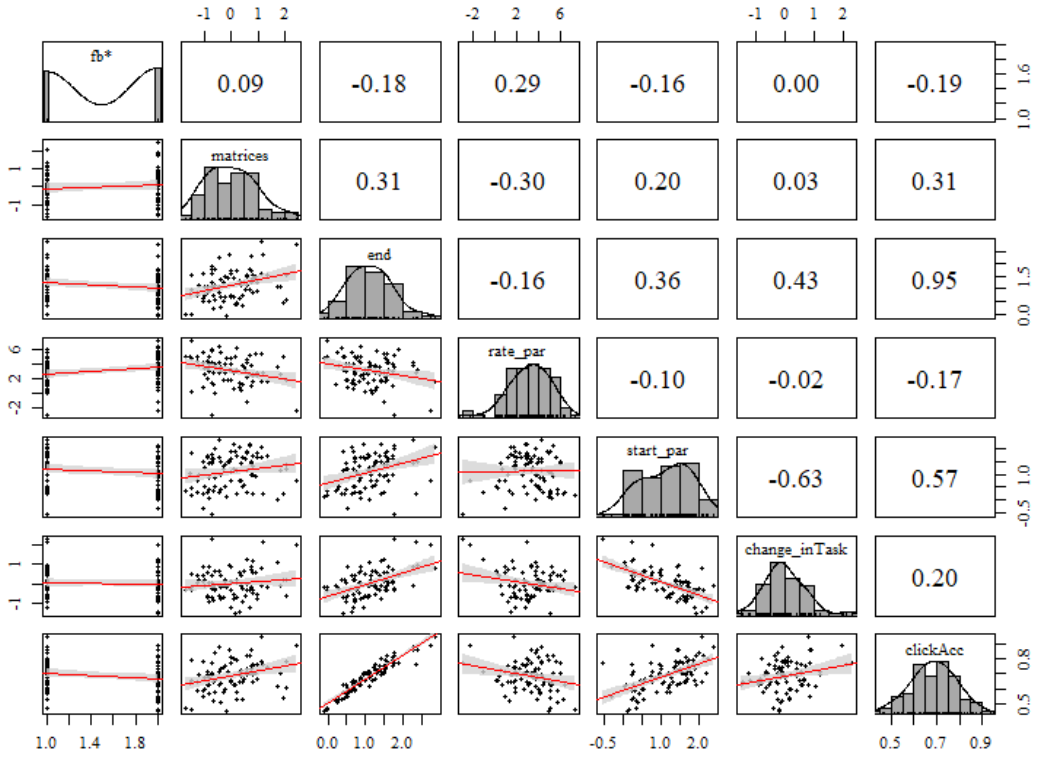


Figure 25. Matrix of scatterplots and correlations for variables of interest. Spearman rank correlations above the diagonal, histograms on the diagonal, and scatterplots with best-fit lines and CI below the diagonal. "matrices" were the separate and combined matrix reasoning measures. "end" and "rate" and "start" correspond to each participant's fit parameters of change in spatial span. "change_inTask" was simply end minus start. "clickAcc" was overall accuracy on spatial span. "fb" indicates presence or absence of feedback.

Several parameters were related to matrix reasoning scores across conditions, and this relation was quite similar to the relation between overall accuracy and matrix scores (r(85) = 0.31 [0.1,0.5], BF_{log2} = 2.93). In addition to ending accuracy (r(85) = 0.33 [0.12,0.52], BF_{log2} = 3.72), rate of change was reliably related to matrix reasoning (r(85) = -0.3 [-0.49,-0.08], BF_{log2} = 2.36). In contrast, both initial accuracy (r(85) = 0.2 [0.01,0.38], BF_{log2} = -0.78) and the magnitude of change (r(85) = 0.07 [-0.15,0.27],

BF_{log2} = -2.81) had negative Bayes Factors, indicating more evidence for the null model than a model supporting a relation between the variables.

Predictiveness time-dependent components of spatial span

Using resampled robust linear models it was evident that between 5% of and 10% of out-of-sample variance in reasoning scores were predicted by both ending accuracy (*b* = 0.55, CI = [0.23,0.87], dR^2_{oos} = 0.1125) and rate of change (*b* = -0.14, CI = [-0.28,-0.015], dR^2_{oos} = 0.0723). Both the effects of ending accuracy and rate of change remain reliable when controlling for the other; in a model including both predictors the dR^2_{oos} of rate was 0.05 while the dR^2_{oos} of ending accuracy was 0.095. These relations were not moderated by feedback; when testing interactions between parameter values and feedback no reliable effects were observed. For this reason, linear models and mediations reported here did not use feedback as a covariate.

Of particular interest was the degree to which links between matrix reasoning and spatial span accuracy were due to learning. Most participants improved over the course of the task (57.5%), and the probability of a participant improving was not linked to feedback condition. Post-hoc analyses indicated that rate of change as well as ending accuracy each continued to explain out-of-sample matrix reasoning scores when tested only in the individuals who improved (rate *b* = -0.14, CI = [-0.32,-0.00013], dR^2_{oos} = 0.088, ending *b* = 0.72, CI = [0.35,1.1], dR^2_{oos} = 0.2262), lending some support to the interpretation that the observed effects were being driven by learning *per se* (see also Appendix 9). Note, however, that including the change in accuracy within a model as a moderator of the effect of rate did not indicate a reliable interaction. Further, including increase or decrease as a moderator of ending accuracy was problematic due to the change score being calculated in part using the ending accuracy. Last, when examining only the participants whose performance decreased over time, there were not reliable correlations between reasoning scores and rate of change (r(35) = -0.18 [-0.45,0.13], BF_{log2} = -1.7) or ending accuracy (r(35) = 0.03 [-0.27,0.39], BF_{log2} = -2.44).

Discussion

Canonical correlations between fluid intelligence measures and working memory task accuracy were replicated. In addition, by utilizing measures of change in working memory accuracy over the course of 68 trials, novel mechanistic insights were observable. While overall accuracy was nearly isomorphic with estimated ending accuracy, including very similar correlations with matrix reasoning scores, time-related model estimates also allowed for uncovering novel relations between reasoning scores and rate of change in spatial span accuracy.

The present results clarified the mechanistic basis of previously observed correlations between working memory measures and fluid intelligence scores. By identifying a near-unity between final working memory accuracy and overall working memory accuracy, overall scores were confirmed to be highly reliable measures of *stable* processes. A complementary finding is supported by the symmetry between correlations of reasoning scores, working memory final accuracy, and overall working memory accuracy. That is, the canonical correlation between fluid intelligence and working memory appears to be due to stable processing characteristics that vary between individuals rather than other possible sources, such as the ability to learn or the ability to immediately perform well on novel tasks or the ability to learn novel tasks.

The present study also provided novel evidence for links between fluid intelligence and the rate of change within working memory tasks. The out-of-sample predictiveness of spatial span rate of change was retained even when controlling for final spatial span accuracy. When controlling for one another, rate of change explained a median 5.0% of variance while final accuracy explained a median 9.5% of variance in held-out samples of matrix reasoning scores (median delta R-squared of 5000 train/test 80/20% divisions). This, along with the lack of a reliable correlation between rate of change and overall spatial span accuracy, further supports the independence of the two links between working memory and fluid intelligence.

There was no reliable support for links between initial working memory ability and fluid intelligence. Correlations between initial spatial span accuracy and reasoning scores indicated a small amount of support for the null hypothesis ($BF_{log2} = -0.78$), with similarly null effects using bootstrapped

robust linear models. The null results cannot support strong claims for a lack of relations between fluid intelligence and initial performance on novel difficult tasks, but it appears clear that canonical relations between working memory and fluid intelligence are unlikely to be primarily due to variation fluid intelligence acting as an index of immediately superior performance (Ackerman & Cianciolo, 2000). One clear implication of this null effect is that, if researchers wish to study the components of working memory performance that are most linked to fluid intelligence, greater numbers of trials may be necessary. Early trials (or all trials in very short batteries of tasks designed to quickly assess cognitive functions) are unlikely to be measuring the processes shared by both fluid intelligence and working memory.

Likewise, links between fluid intelligence measures and the magnitude of learning were not supported in the current data. Compelling evidence was observed for the null effect (i.e., no relations between fluid intelligence and magnitude of learning; $BF_{log2} = -2.81$). The separability of rate and magnitudes of learning are necessarily dependent on the learning domain and the timescale of change (see Study 3.3). Early in learning rate and magnitude are largely conflated, with distinctions only possible on timescales wherein asymptotic change is possible. While the complications due to domain and timescale preclude direct comparisons between the current results and other reports of links between magnitude of learning and fluid intelligence (e.g., Ren et al., 2015), it is still notable that improvements on working memory tasks are possible for many hundreds or even thousands of trials (Jaeggi et al., 2010), indicating that the change observed in the working memory task here were not simply due to a rapidly-asymptoting nature of the within-domain learning. Indeed, given the relatively short timescale of the task, it is notable that the magnitude and rate of change were convincingly uncorrelated ($BF_{log2} = -2.46$).

Relations between fluid intelligence and working memory may have multiple mechanistically distinct bases. Both rate of change and asymptote of spatial span accuracy were independently predictive of fluid intelligence. Conventional approaches time-dependent change (i.e., ignoring it) would have conflated the two sets of relations. Further, studies implementing working memory tasks of different trial numbers or difficulties may have inadvertently been disproportionately measuring one or the other

underlying processes giving rise to working memory performance (i.e., time-dependent vs. asymptotic aspects), and thereby introducing preventable heterogeneity and inconsistency into the study of relations between fluid intelligence and working memory.

Chapter 6. Pragmatics of a continuous-time psychology

The systematic study of the nonlinear dynamics of learning and forgetting is many decades old (Crossman, 1959; Snoddy, 1926). Quantitatively rigorous approaches to the study of learning functions have thus not been held back by theory alone. Further, modern computing power has allowed for the realization of algorithms and software to solve complex and cumbersome that would have been intractable in the past. A brief historical review of learning curve analysis demonstrates this point. Early investigations of learning curves, and even continuing into the era of modern computers, utilized simplistic methods of drawing linear trends through log-transformed time (Crossman, 1959; A. Newell & Rosenbloom, 1981; Snoddy, 1926). With the widespread availability of powerful computing and advanced software more complex nonlinear models have become possible (Dosher & Lu, 2007; Heathcote et al., 2000). Still, the implementation of by-trial modeling has remained limited (cf. Cochrane et al., 2019; Karl M. Newell et al., 2009; Reddy et al., 2018). Applications of by-trial nonlinear modeling of non-Gaussian distributions is exceedingly rare, in no small part due to challenges with model specification and estimation.

In modeling contexts with no closed-form solution (e.g., many nonlinear models) two broad categories of model fitting are commonly implemented, maximum likelihood (ML) and Bayesian sampling. The first seeks to find the parameter combination that minimizes a given error function (i.e., maximizes the corresponding likelihood function), while the second seeks to characterize a multivariate parameter space using Monte Carlo techniques. The relative merits and implementational differences between ML and Bayesian methods are beyond the scope of the current document. However, it is useful to note that previous analyses of learning curves have overwhelmingly used ML fitting of group-level data (A. Newell & Rosenbloom, 1981; Karl M. Newell et al., 2009) or of individual-level data (Cochrane et al., 2019; Dosher & Lu, 2007; Heathcote et al., 2000). In contrast, in this dissertation several sections have included Bayesian models with hierarchical regression structures allowing for simultaneous fit of individual-level learning curves and group-level parameter distributions. This straightforward extension

of Bayesian methods to nonlinear hierarchical models, along with the information about models available as a result of the fitting method (e.g., full marginal and joint parameter distributions), makes Bayesian fitting likely to be the gold standard in the analysis of learning studies.

In any context of model fitting and the corresponding theory-driven interpretation of data, model specification is a matter of parsimony and pragmatics (among other factors). Increased model complexity increases both the likelihood of overfitting (i.e., interpreting idiosyncratic patterns as if they were generalizable) as well as the sheer difficulty in implementing model fits. The issue of overfitting may be addressed using statistical model comparison techniques, by increasing data quantity, by utilizing converging evidence from various sources of data, or other methods; this topic is outside the scope of the current chapter. The second issue, that of practical challenges, will be considered further here. From a practical standpoint Bayesian hierarchical nonlinear model fitting has serious problems. Models take an inordinate amount of time to fit, with some of the models reported in this manuscript having run for several weeks, and prior knowledge about the nature of the to-be-fit model is often crucial.

The ideal analytical methods are necessarily guided by knowledge about the generative process of, and the constraints on, data. This includes information about expected distributions (e.g., response times and accuracies are non-Gaussian) and speed of possible process change (e.g., questions regarding whether theoretically interesting change is possible in 2 trials or 500 trials). Full integration of distributional information is most naturally implemented in fully Bayesian models, which is the reason why many models described in this manuscript have been fit using nonlinear hierarchical regression parameterized and modeled using fully Bayesian methods. This included, variously, models specified using **JAGS** and Gibbs sampling (Plummer, 2003) or in the **brms** front-end for **R** to the **Stan** Hamiltonian Monte Carlo software (Bürkner, 2017). However, these analytical methods are slow and they are complicated. While the core message of this manuscript (i.e., that systematic time-dependent dynamics are central to psychological theories) is intended to be widely applicable and approachable, there is little promise in asking researchers to learn a new system of statistics and wait many days to see if they have succeeded. Instead, a relatively simple and fast implementation of time-evolving models is

warranted. In order to make time-evolving analyses of data practical for the everyday psychology researcher, I have written an open source R package to automate the analysis process for many common types of behavioral data. This package (**TEfits**) that can be installed in R with two simple commands (*install.packages('devtools'); devtools::install_github('akcochrane/TEfits'*)).

Statistical package: TEfits (Cochrane, 2020)

This statistical package streamlines and automates many tasks associated with modeling learning and with ML fitting of nonlinear regression more generally. Implementation uses a minimum of dependencies on other statistical packages. The package vignettes and other documentation explain the use, methods, and assumptions of **TEfits**; the focus later in this chapter is on performance. Specifically, in the upcoming section I systematically compare **TEfits** performance to the performance of hierarchical Bayesian model fitting using **brms**. In addition to the current performance analysis and the documentation provided with the package on Github, the following manuscript outlines the broad motivation, purpose, and implementation of the package. The manuscript has been published at the Journal of Open Source Software (open review: https://github.com/openjournals/joss-reviews/issues/2535; see Appendix 10).

Simulations comparing TEfits to mixed-effects Bayesian models

Methods

Above and beyond the clear *a priori* benefits and challenges of Bayesian or ML modeling discussed earlier in this chapter, an empirical comparison of fitting methods was warranted to make the practical choice more concrete. Specifically, by simulating data, fitting that data using each method, and comparing the quality of fits, I demonstrate the implications of various analytical and methodological choices.

In this chapter I have recapitulated the methods of Chapter 2. In the earlier chapter the emphasis was on the differential effects of time-sensitive or time-insensitive analyses in relation to variations in the underlying generative process of the data. In this chapter the variations in generative process will be

randomly generated and the interaction will be analyzed between methodological choices and analytical choices (i.e., hierarchical Bayesian modeling vs. **TEfits**)

Data generation

Simulated datasets varied on two variables under experimental control. In most experiments methodological choices by the experimenter cause there to be some measurement for each experimental trial as well as some number of observed trials. Each trial measurement could have some amount of variability, corresponding to information per data point, such as binary (e.g., present/absent or N-alternative-forced-choice), percent correct (e.g., 5 out of 9 correct on a memory span trial leading to a 55.6% trial accuracy), or a combination of categorical and response time information. The variation in possible outcomes per trial is directly proportional to the degree of precision regarding the estimation of underlying expected values (e.g., with increasing number of binary values in a binomial distribution, the standard deviation of the normalized [zero-to-one-scaled] distribution decreases proportionally to the square root of the number of values). As in Chapter 2, in this chapter I have simulated a memory task with varying possible numbers of correct answers per trial.

Data was simulated in two dimensions on spectra from “coarse” (i.e., few observations per trial and/or low number of trials) to “fine” (i.e., many observations per trial and/or many trials). Although the data could have analogously been simulated from other distributions (e.g., Gaussian), the binomial distribution used here provides a direct comparison to practical experimental psychology methods while also being analogous to other distributions. Data was simulated using a “participant number” of 200, a reasonable number for behavioral experiments while also being large enough to approximate asymptotic properties of simulations. Methodological variation included n-per-trial ranging from 1 to 6 and trial-number ranged from 25 to 200. Asymptotically good fits are approached toward the higher range of each of these dimensions.

Each simulated participant’s data consisted of a vector of random data sampled from a binomial distribution with a changing expected value (i.e., binomial probability of a “success”). The expected value for each participant changed as a deterministic exponential function of trial number (see Appendix 1). The

starting values (`0 < start < .5`), asymptotic values (`0.5 < asymptote < 1.0`), and rates [shapes] (`2 < rate < log~2~(max_trial_number)`) of each deterministic exponential function were sampled uniformly from the given ranges.

Modeling methods

TEfits models were fit to each simulated dataset (i.e., combination of n-per-trial and trial-number) using defaults of the **TEfitAll** function. Briefly, this fits the nonlinear function `asymptote + (start - asymptote)*exp((1-trialNumber)/2^rate)`, which is identical to the generating function in order to allow for parameter comparisons with the true values. A Bernoulli error function was used (i.e., negative log Bernoulli likelihood). An individual model was fit to each simulated participant's data.

Bayesian hierarchical fitting using **brms** proceeded using the same nonlinear function as **TEfits**. Each parameter was modeled as a distribution of participant-level random effects, which were then extracted and compared to the true generated values. Fixed effects of start and asymptote parameters used flat beta-distributed priors, while the rate parameter used a Gaussian prior centered at 4 with a standard deviation of 3. All other priors used **brms** defaults.

Comparison methods

Two indices were used to compare models' goodness-of-fit. As the true generating values of each parameter were known, fit parameters could be compared to the true values using the root mean squared error (RMSE), and product-moment correlations. RMSE provided measures of distance between fit values and true values while correlation provided measures of the relative relatedness of true and fit values. Each of these indices was calculated for each dataset's **brms** fit and **TEfits** fit.

In plots of all comparisons, colors are scaled such that the direction in which "**TEfits** is better" is red; this is when contrasting TEfits with a relative criterion (e.g., higher RMSE compared to BRMS is worse) or an absolute criterion (e.g., higher RMSE is worse; see Appendix 11).

Results

Comparisons of RMSE

Figures 26, 27, and 28 show the RMSE of **brms** models minus RMSE of **TEfits** models. In every simulation, across data information per data point (i.e., *n per trial*) and trial number, the hierarchical Bayesian models outperformed the by-participant maximum likelihood models. Methodological changes in each dimension are able to mitigate the difference between model types. **TEfits** does not approach the accuracy of **brms** in estimating rate or asymptote unless both dimensions of data information are increased. In contrast, only increases in per-trial information appears necessary to mitigate the differences between models' estimation of starting parameters.

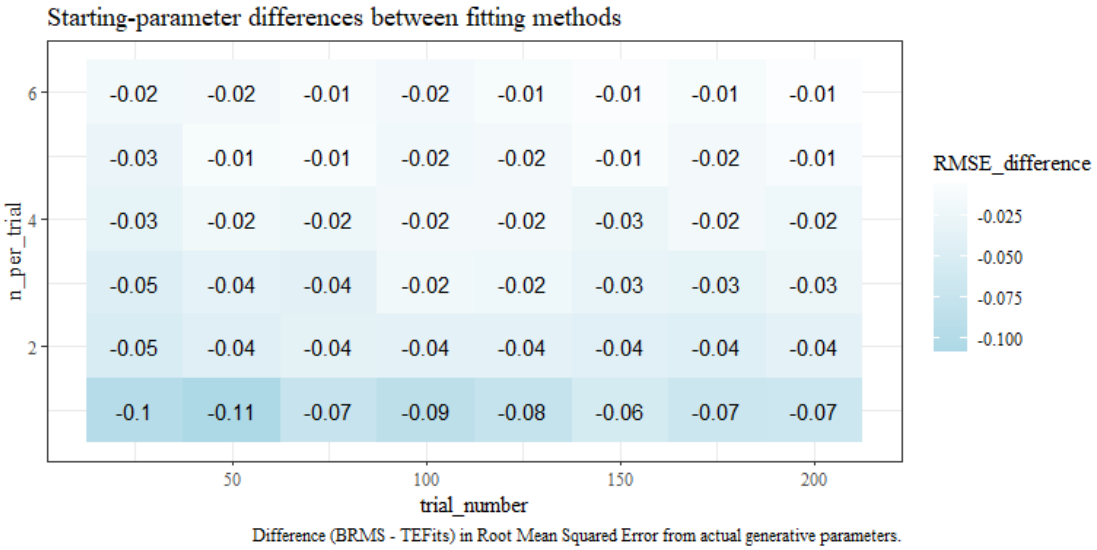


Figure 26. Differences between TEfits and Bayesian multilevel nonlinear model in RMSE of fit starting-parameter values. Larger values (i.e., less negative) indicate better performance by TEfits. Starting parameters are estimated well when each trial has more information (i.e., greater n_per_trial) but the quality of estimates is less influenced by number of trials.

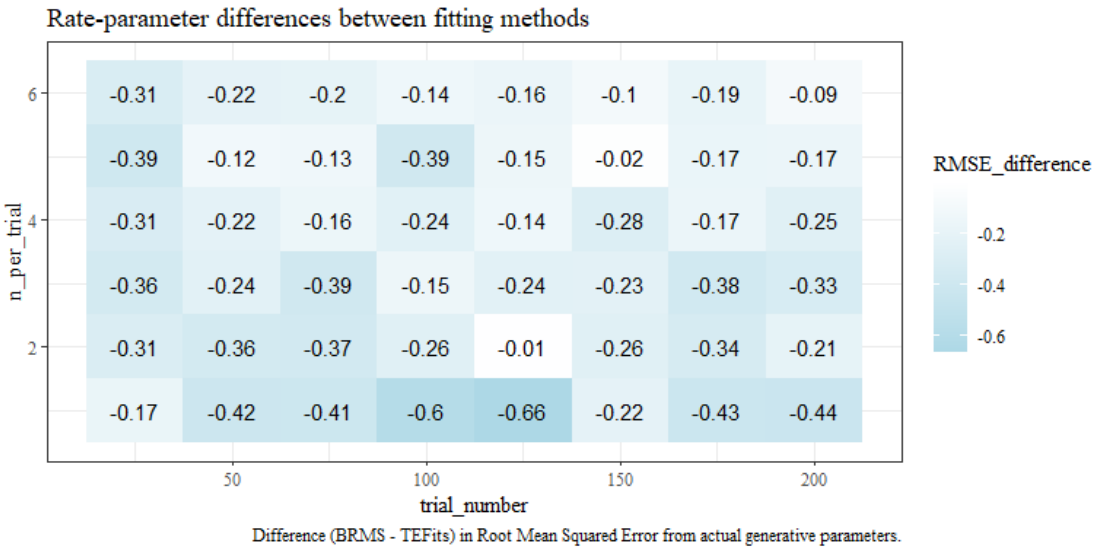


Figure 27. Differences between TEfits and brms Bayesian multilevel nonlinear modeling in RMSE of fit rate-parameter values. Larger values (i.e., less negative) indicate better performance by TEfits. TEfits performs poorly with less information in either dimension (i.e., less information per trial, n_per_trial, or fewer trials, smaller trial_number). Only with large amounts of information per trial and large numbers of trials does the accuracy of TEfits approach that of brms.

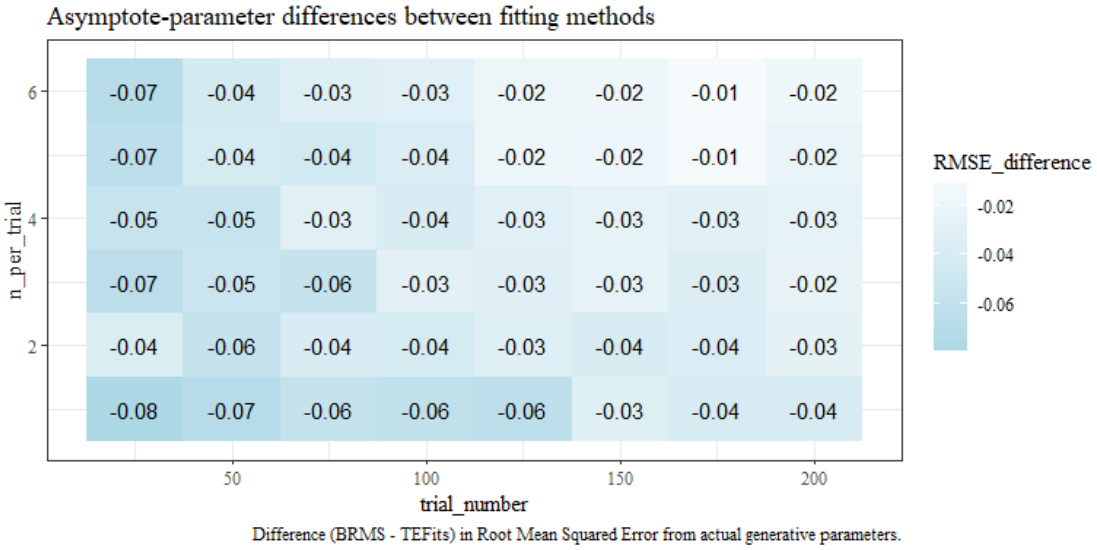


Figure 28. Differences between TEfits and brms Bayesian multilevel nonlinear modeling in RMSE of fit asymptote-parameter values. Larger values (i.e., less negative) indicate better performance by TEfits.

Quality of estimated asymptotes varies primarily by the number of trials utilized, with uniformly poorer relative performance of TEfits at smaller trial numbers.

Comparisons of correlations

Figures 29, 30, and 31 show the correlations between estimated parameters and true parameters of **brms** models minus the corresponding correlations of **TEfits** models. In almost all simulations, across both dimensions of methodological choices, the hierarchical Bayesian models outperformed the by-participant maximum likelihood models. Asymptote and rate parameters appeared to only be benefited by increasing trial numbers, which clearly indicates the difficulty of estimating the shape or end-point extrapolated from a small amount of data. In contrast, starting parameters are equally well estimated by both methods when *n per trial* is high and the number of trials has little influence on the differences between methods.

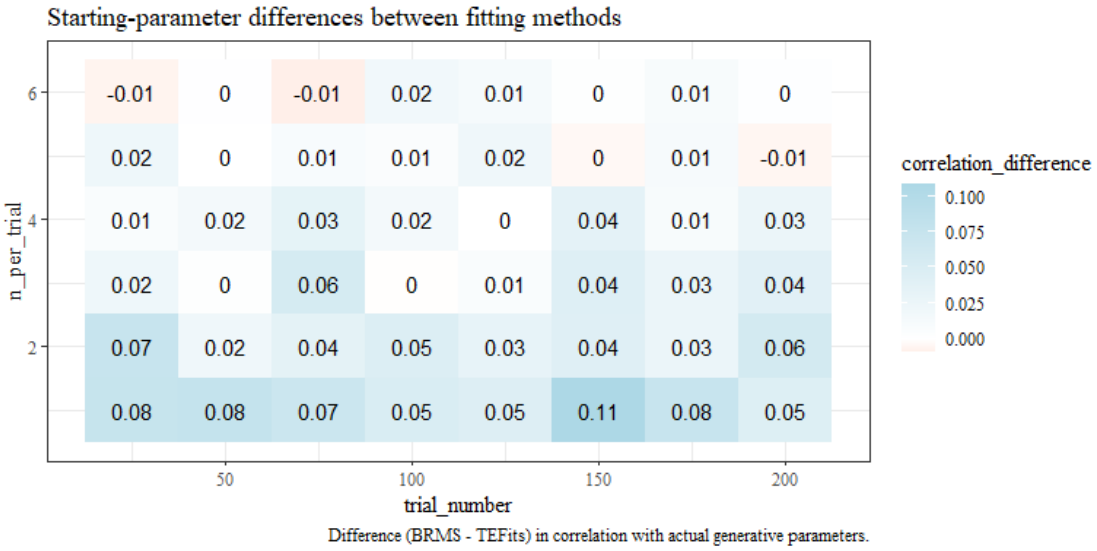


Figure 29. Differences between TEfits and brms Bayesian multilevel nonlinear modeling in correlations between true and estimated parameter values. Smaller values (i.e., closer to zero) indicate better performance by TEfits. TEfits performs just as well as brms with a large amount of information per trial (n_per_trial), and TEfits is largely unaffected by variations in trial number.

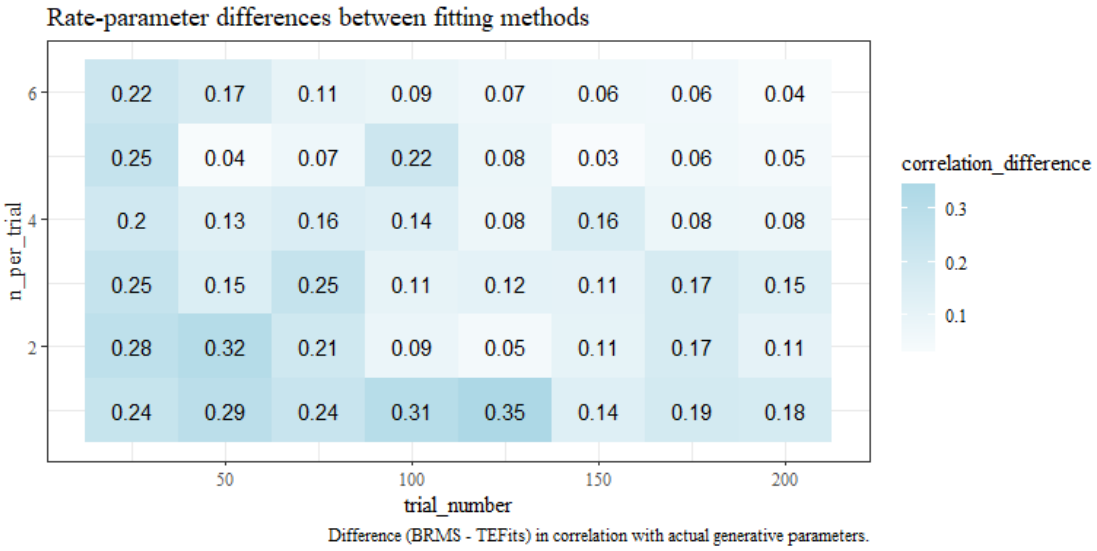


Figure 30. Differences between TEfits and brms Bayesian multilevel nonlinear modeling in correlations between true and estimated parameter values. Smaller values (i.e., closer to zero) indicate better performance by TEfits. Greater information per trial (i.e., n_per_trial) or numbers of trials each benefit TEfits independently, but a combination of both is needed to approach the quality of brms fits.

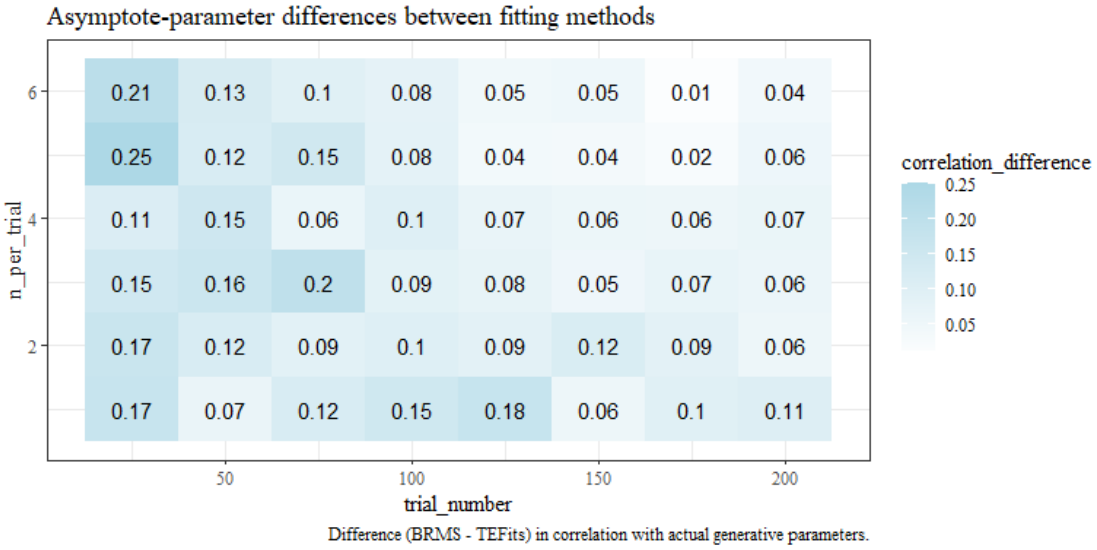


Figure 31. Differences between TEfits and brms Bayesian multilevel nonlinear modeling in correlations between true and estimated parameter values. Smaller values (i.e., closer to zero) indicate better

performance by TEfits. Quality of TEfits are worst with a large amount of information per trial (i.e., n_per_trial) and a small trial number. With increasing trial numbers, the quality of TEfits approaches that of brms.

Results summary

It is clear that the multilevel Bayesian **brms** fitting method is consistently superior to the **TEfits** method, with the latter method establishing a compelling "gold standard."" Additionally, these differences can be easily mitigated through methodological choices that increase trial number and per-trial data information. Nonetheless, in absolute terms **TEfits** provides acceptably accurate estimates of true parameters; the current results also provide approximate guidance regarding links between methodological choices (e.g., number of trials) and the precision of time-evolving parameter estimates.

Discussion of pragmatics of continuous-time implementations

TEfits is not intended to replace analyses that are able to incorporate fuller knowledge of the structure and generative process underlying data (e.g., the hierarchical nonlinear Bayesian approach using **brms** that has frequently been used in this manuscript). Instead, my goal in writing **TEfits** was to simplify the implementation of a time-centered approach to understanding behavioral data through a combination of interpretable parameterizations, repeated function-optimization runs (with or without resampling techniques), and sensible defaults for many types of behavioral data.

Simplification clearly has drawbacks, however, as was evident in the decrease in accuracy of **TEfits** relative to **brms** in simulated data. The difference between approaches is somewhat mitigated by methodological choices such as improved data fidelity (e.g., utilizing continuous rather than binary data) or by increasing numbers of trials. Indeed, the importance of the number of trials cannot be overstated, as trial number constrains the range of timescales of valid estimation and ultimately the interpretability of values. Unless there are enough trials to model an inflection, then any learning curve will imitate a linear function, and asymptote parameters will be uninterpretable.

One methodological consideration that has not been directly considered here is that of explicitly non-*iid* data. Study 5.2 briefly addressed the difficulty in modeling change when trials are explicitly non-interchangeable (e.g., pattern-completion problems). Many other methodological approaches likewise involve trials of non-interchangeable measurements of behavior that are explicitly non-*iid* (even conditional on a time-evolving trend). A common experimental approach, that of adaptive stimulus presentation, is likely to preclude time-dependent modeling. The conflation of stimulus difficulty, time, and measured performance simply reduces the information available to estimate changes in underlying psychological processes. In order for learning to be assessed, experiments must be designed to allow early performance to be compared to late performance.

A core emphasis of this dissertation, and of **TEfits**, is the recognition of the importance of the dimension of time. There is no single approach that can always effectively address the extent to which data is or is not *iid*. Instead **TEfits** implements several such tests. Simple correlations of data with trial number provides a first approximation. One of the simplest to interpret summaries of a **TEfits** model is the comparison of the extent to which data is conditionally (on a fit model) or unconditionally correlated with trial number. Along similar lines, **TEfits** incorporates a default Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test of level stationarity for the residuals of both the time-invariant and time-evolving models. To the extent that the residuals of the stable model exhibit nonstationarity while the residuals of the time-evolving model do not reject the null hypothesis of the KPSS test, this provides evidence for the relative benefit of a continuous-time model. Other summaries of time-evolving fits to data involve comparison of the Bayesian Information Criterion (BIC) between null and time-continuous models, calculating the out-of-sample reduction of error in random subsamples, or fitting many resampled models and calculating the percent of these models that predict increasing or decreasing fits to data. Clearly there are many approaches to assessing the appropriateness of time-evolving models, and each of the approaches should be considered in the context of the others. Each one indicates the extent to which data systematically changes over the course of the dimension of time.

I have also implemented an approach that wholly emphasizes ease-of-use by implementing the syntax of the popular multilevel model R package **lme4** (Bates et al., 2015). Although this last method has a restricted set of possible model specifications and takes certain shortcuts (i.e., pre-estimating the shape or "rate" of learning before full model fitting), it thereby removes almost all barriers to approaching behavior as continuously evolving over time. Any user of **lme4** should be able to use **TEfits** models without any additional skill needed.

TEfits is in continuous development. Many features are planned, including additional distributions, fuller documentation, simplified syntax, and fully featured hierarchical Bayesian model refitting. Additional distributions will include the Wiener diffusion process for modeling continuous changes in DDM parameters. Documentation will include full vignettes for all major methods implemented in **TEfits**. Syntax will optionally be more symbolic, which should improve the ability of new users to learn how to use the package. Lastly, and most importantly, the currently implemented automatic refitting of **TEfits** models using Bayesian hierarchical methods should allow the performance decrements reported in this chapter to be mitigated through the implementation of both modeling approaches.

Chapter 7. Concluding remarks

The call for considering experience-dependent changes on a trial-by-trial basis has a long history (e.g., Harlow, 1949). Likewise, the argument for greater consideration of intrapersonal change in psychological processes is far from new (Heathcote et al., 2000; Molenaar, 2004; Thelen & Smith, 1994). In this manuscript, I have demonstrated the theoretical and empirical importance of unifying these perspectives. I have used nonlinear computational methods to extend common analytical approaches, such as fitting psychometric functions, with inferences regarding intra-personal change on the timescale of individual learning events.

In both simulations and empirical studies of perceptual learning, my colleagues and I demonstrated the benefits, in terms of model goodness-of-fit as well as theoretical inferences, of by-trial approaches to learning (Chapter 2; Studies 3.1 & 3.2). I next provided evidence from two perceptual learning experiments that functions of change coming from an exponential family were best suited to characterize by-trial changes due to learning (Study 3.3). In subsequent work considering behavioral experiments that were nominally not directly concerned with learning, I showed applications of a learning-centered perspective in two response competition tasks (Studies 4.1, 4.2 , & 4.3a-c). I then demonstrated that theoretical questions of individual differences, both in putative learning tasks and non-learning task, are better addressed by eschewing assumptions of process stationarity and instead using time-evolving indices of task performance (Studies 5.1 & 5.2). Lastly, I introduced the statistical package I have written to allow behavioral researchers to integrate inferences regarding trial-timescale changes into conventional analytical environments (Chapter 6).

Each of the studies reported here should motivate further examination of time-dependent processes. For example, evidence for learning to learn reported in Study 3.2 raised the possibility that previously observed patterns of generalization (or lack thereof) in many fields of learning may have been mischaracterized due to assumptions of stationarity within tests of generalization. Chapter 3 additionally corroborated the mounting evidence against the use of power functions to interpret processes of learning

in Study 3.3, and instead showed that tasks may differ in their learning functions despite belonging to the same functional family. As in Chapter 3, likewise Chapters 4 and 5 should generate as many questions as they do answers.

One feature of the studies in this dissertation cannot be overlooked: Despite my repeated calls for testing assumptions of stationarity, I just as frequently make the assumption of stationarity in many measures. In Study 5.2 I assumed matrix reasoning scores were stationary over time because accuracy on different items could not be assumed to be *iid* (either unconditionally or conditioned on time). The dimension of difficulty would need to be empirically established across matrix problems prior to continuous-time modeling. In Study 5.1 my colleagues and I could have decomposed every behavioral variable into time-dependent components, but this would have left us with an intractable number of predictors with an unknown underlying dimensionality. Instead we computed aggregated measures of performance while accepting the possible conflation of underlying processes.

Due to the increased dimensionality of data that arises from decomposing time-dependent components of performance, one particularly important step in shifting psychological science to a more person-centered regime is the use of multivariate time series (Molenaar, 2004). Each of the experiments reported in this manuscript emphasizes the theoretical and empirical implications of treating individuals' psychological processes as continuously unfolding over time. It goes without saying that, to take this proposition to its logical conclusion, multiple dimensions of psychological processes would need to be concurrently modeled. Multivariate time series have been advanced on short timescales, such as coupled EEG and fMRI measures (Jorge et al., 2014; Lemieux et al., 1997), and at long timescales of days to years (Deboeck et al., 2009; Hertzog et al., 2008). Coupled dynamics on timescales of minutes to hours could be considered the most common topic of experimental psychology (i.e., assessing effects and interactions associated with manipulations), yet there has been a striking lack of corresponding multivariate approaches that treat changes in behavior as truly continuous. As described in the previous paragraph, even in this dissertation I have largely considered only univariate changes (i.e., in various individual measures of "performance").

Interactions between organism and environment lead to systematic changes within the organism, and this manuscript has been a call for a greater consideration of the change occurring within psychological experiments. Most psychological processes are, to some extent, influenced by learning processes. It should thus be worthwhile to integrate a continuous-time, and often learning-centered, perspective into perception, attention, memory, and diverse other domains.

References

Ackerman, P. L., Beier, M. E., & Boyle, M. O. (2005). Working memory and intelligence: The same or different constructs? *Psychological Bulletin*, *131*(1), 30–60. https://doi.org/10.1037/0033-2909.131.1.30

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied*, *6*(4), 259–290. https://doi.org/10.1037/1076-898X.6.4.259

Ackerman, P. L., Kanfer, R., & Goff, M. (1995). Cognitive and noncognitive determinants and consequences of complex skill acquisition. *Journal of Experimental Psychology: Applied*, *1*(4), 270–304. https://doi.org/10.1037/1076-898X.1.4.270

Adams, J. A. (1952). Warm-up Decrement in Performance on the Pursuit-Rotor. *The American Journal of Psychology*, *65*(3), 404. https://doi.org/10.2307/1418761

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences*, *90*(12), 5718–5722. https://doi.org/10.1073/pnas.90.12.5718

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*(6631), 401–406. https://doi.org/10.1038/387401a0

Ahissar, M., & Hochstein, S. (2000). The spread of attention and learning in feature search: Effects of target distribution and task difficulty. *Vision Research*, *40*(10–12), 1349–1364. https://doi.org/10.1016/S0042-6989(00)00002-X

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences*, *8*(10), 457–464. https://doi.org/10.1016/j.tics.2004.08.011

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1515), 285–299. https://doi.org/10.1098/rstb.2008.0253

Alloway, T. P., & Alloway, R. G. (2013). Working memory across the lifespan: A cross-sectional approach. *Journal of Cognitive Psychology*, *25*(1), 84–93. https://doi.org/10.1080/20445911.2012.748027

Anderson, J. R., Fincham, J. M., & Douglass, S. (1999). Practice and retention: A unifying analysis. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *25*(5), 1120–1136.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Ball, K., & Sekuler, R. (1987). Direction-specific improvement in motion discrimination. *Vision Research*, *27*(6), 953–965. https://doi.org/10.1016/0042-6989(87)90011-3

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bavelier, D., Green, C. S., Pouget, A., & Schrater, P. (2012). Brain Plasticity Through the Life Span: Learning to Learn and Action Video Games. *Annual Review of Neuroscience*, *35*(1), 391–416. https://doi.org/10.1146/annurev-neuro-060909-152832

Bavelier, D., Levi, D. M., Li, R. W., Dan, Y., & Hensch, T. K. (2010). Removing Brakes on Adult Brain Plasticity: From Molecular to Behavioral Interventions. *Journal of Neuroscience*, *30*(45), 14964–14971. https://doi.org/10.1523/JNEUROSCI.4812-10.2010

Bejjanki, V. R., Beck, J. M., Lu, Z.-L., & Pouget, A. (2011). Perceptual learning as improved probabilistic inference in early sensory areas. *Nature Neuroscience*, *14*(5), 642–648. https://doi.org/10.1038/nn.2796

Bejjanki, V. R., Zhang, R., Li, R., Pouget, A., Green, C. S., Lu, Z.-L., & Bavelier, D. (2014). Action video game play facilitates the development of better perceptual templates. *Proceedings of the National Academy of Sciences*, *111*(47), 16961–16966. https://doi.org/10.1073/pnas.1417056111

Berch, D. B., Krikorian, R., & Huha, E. M. (1998). The Corsi Block-Tapping Task: Methodological and Theoretical Considerations. *Brain and Cognition*, *38*(3), 317–338. https://doi.org/10.1006/brcg.1998.1039

Binet, A., & Simon, Th. (1916). The development of intelligence in the child. In E. S. Kite (Trans.), *The development of intelligence in children (The Binet-Simon Scale)*. (pp. 182–273). Williams & Wilkins Co. http://content.apa.org/books/11069-004

Bornstein, M. H., Hahn, C.-S., Bell, C., Haynes, O. M., Slater, A., Golding, J., Wolke, D., & the ALSPAC Study Team. (2006). Stability in Cognition Across Early Childhood: A Developmental Cascade. *Psychological Science*, *17*(2), 151–158. https://doi.org/10.1111/j.1467-9280.2006.01678.x

Brooks, V., Hilperath, F., Brooks, M., Ross, H. G., & Freund, H. J. (1995). Learning “what” and “how” in a human motor task. *Learning & Memory*, *2*(5), 225–242. https://doi.org/10.1101/lm.2.5.225

Burke, L. A., & Hutchins, H. M. (2007). Training Transfer: An Integrative Literature Review. *Human Resource Development Review*, *6*(3), 263–296. https://doi.org/10.1177/1534484307303035

Bürkner, P.-C. (2017). **brms**: An *R* Package for Bayesian Multilevel Models Using *Stan*. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Byers, A., & Serences, J. T. (2012). Exploring the relationship between perceptual learning and top-down attentional control. *Vision Research*, *74*, 30–39. https://doi.org/10.1016/j.visres.2012.07.008

Casasanto, D., & Lupyan, G. (2015). All concepts are ad hoc concepts. In E. Margolis & S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 543–566). MIT Press.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. *Journal of Educational Psychology*, *54*(1), 1–22. https://doi.org/10.1037/h0046743

Cochrane, A., Cui, L., Hubbard, E. M., & Green, C. S. (2019). "Approximate number system" training: A perceptual learning approach. *Attention, Perception, & Psychophysics*, *81*(3), 621–636. https://doi.org/10.3758/s13414-018-01636-w

Cochrane, A., Simmering, V., & Green, C. S. (2020). Load effects in attention: Comparing tasks and age groups. *Attention, Perception, & Psychophysics*, *82*(6). https://doi.org/10.3758/s13414-020-02055-6

Cochrane, A., Simmering, V. R., Austerweil, J. L., & Green, C. S. (2018). Rapid Learning in Early Attentional Processing: Bayesian Estimation of Trial-by-Trial Updating. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *40*, 232–237.

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology*, *85*(5), 678–707. https://doi.org/10.1037/0021-9010.85.5.678

Constantinidis, C., & Klingberg, T. (2016). The neuroscience of working memory capacity and training. *Nature Reviews Neuroscience*, *17*(7), 438–449. https://doi.org/10.1038/nrn.2016.43

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Cowan, N., & Alloway, T. P. (2009). The development of working memory. In M. Courage & N. Cowan (Eds.), *The development of memory in infancy and childhood* (pp. 303–342). Psychology Press.

Cox, W. T. (2015). *Multiple determinants of prejudicial and nonprejudicial behavior* [Doctoral Dissertation]. University of Wisconsin - Madison.

Crossman, E. R. F. W. (1959). A THEORY OF THE ACQUISITION OF SPEED-SKILL. *Ergonomics*, *2*(2), 153–166. https://doi.org/10.1080/00140135908930419

Deboeck, P. R., Montpetit, M. A., Bergeman, C. S., & Boker, S. M. (2009). Using derivative estimates to describe intraindividual variability at multiple time scales. *Psychological Methods*, *14*(4), 367–386. https://doi.org/10.1037/a0016622

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Dosher, B. A., Jeter, P., Liu, J., & Lu, Z.-L. (2013). An integrated reweighting theory of perceptual learning. *Proceedings of the National Academy of Sciences*, *110*(33), 13678–13683. https://doi.org/10.1073/pnas.1312552110

Dosher, B. A., & Lu, Z.-L. (2007). The Functional Form of Performance Improvements in Perceptual Learning: Learning Rates and Transfer. *Psychological Science*, *18*(6), 531–539. https://doi.org/10.1111/j.1467-9280.2007.01934.x

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149.

Fahle, M. (2005). Perceptual learning: Specificity versus generalization. *Current Opinion in Neurobiology*, *15*(2), 154–160. https://doi.org/10.1016/j.conb.2005.03.010

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*(3), 340–347. https://doi.org/10.1162/089892902317361886

Fry, A. F., & Hale, S. (1996). Processing speed, working memory, and fluid intelligence: Evidence for a developmental cascade. *Psychological Science*, *7*(4), 237–241. https://doi.org/10.1111/j.1467-9280.1996.tb00366.x

Fukuda, K., Vogel, E., Mayr, U., & Awh, E. (2010). Quantity, not quality: The relationship between fluid intelligence and working memory capacity. *Psychonomic Bulletin & Review*, *17*(5), 673–679. https://doi.org/10.3758/17.5.673

Gallistel, C. R., Fairhurst, S., & Balsam, P. (2004). The learning curve: Implications of a quantitative analysis. *Proceedings of the National Academy of Sciences*, *101*(36), 13124–13131. https://doi.org/10.1073/pnas.0404965101

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology. General*, *121*(4), 480–506.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: The implicit association test. *Journal of Personality and Social Psychology*, *74*(6), 1464.

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review*, *56*(1), 51–65.

Harrison, T. L., Shipstead, Z., & Engle, R. W. (2015). Why is working memory capacity related to matrix reasoning tasks? *Memory & Cognition*, *43*(3), 389–396. https://doi.org/10.3758/s13421-014-0473-3

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207.

Hertzog, C., Kramer, A. F., Wilson, R. S., & Lindenberger, U. (2008). Enrichment Effects on Adult Cognitive Development: Can the Functional Capacity of Older Adults Be Preserved and Enhanced? *Psychological Science in the Public Interest*, *9*(1), 1–65. https://doi.org/10.1111/j.1539-6053.2009.01034.x

Horn, J. L., & Cattell, R. B. (1966). Refinement and test of the theory of fluid and crystallized general intelligences. *Journal of Educational Psychology*, *57*(5), 253–270.

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. *Memory & Cognition*, *42*(3), 464–480. https://doi.org/10.3758/s13421-013-0364-z

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y.-F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning—Implications for training and transfer. *Intelligence*, *38*(6), 625–635. https://doi.org/10.1016/j.intell.2010.09.001

Jeter, P. E., Dosher, B. A., Petrov, A., & Lu, Z.-L. (2009). Task precision at transfer determines specificity of perceptual learning. *Journal of Vision*, *9*(3), 1–1. https://doi.org/10.1167/9.3.1

Jorge, J., van der Zwaag, W., & Figueiredo, P. (2014). EEG–fMRI integration for the study of human brain function. *NeuroImage*, *102*, 24–34. https://doi.org/10.1016/j.neuroimage.2013.05.114

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: A latent-variable approach to verbal and visuospatial memory span and reasoning. *Journal of Experimental Psychology. General*, *133*(2), 189–217. https://doi.org/10.1037/0096-3445.133.2.189

Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, *2*(10), 389–398.

Kattner, F., Cochrane, A., Cox, C. R., Gorman, T. E., & Green, C. S. (2017). Perceptual Learning Generalization from Sequential Perceptual Training as a Change in Learning Rate. *Current Biology*, *27*(6), 840–846. https://doi.org/10.1016/j.cub.2017.01.046

Kattner, F., Cochrane, A., & Green, C. S. (2017). Trial-dependent psychometric functions accounting for perceptual learning in 2-AFC discrimination tasks. *Journal of Vision*, *17*(11), 3. https://doi.org/10.1167/17.11.3

Kersten, D., Mamassian, P., & Yuille, A. (2004). Object perception as Bayesian inference. *Annual Review of Psychology*, *55*, 271–304. https://doi.org/10.1146/annurev.psych.55.090902.142005

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics*, *63*(8), 1421–1455. https://doi.org/10.3758/BF03194552

Kyllonen, P. C., & Christal, R. E. (1990). Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, *14*(4), 389–433. https://doi.org/10.1016/S0160-2896(05)80012-1

Law, C.-T., & Gold, J. I. (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature Neuroscience*, *12*(5), 655–663. https://doi.org/10.1038/nn.2304

Law, C.-T., & Gold, J. I. (2010). Shared mechanisms of perceptual learning and decision making. *Topics in Cognitive Science*, *2*(2), 226–238. https://doi.org/10.1111/j.1756-8765.2009.01044.x

Lehle, C., & Hübner, R. (2008). On-the-fly adaptation of selectivity in the flanker task. *Psychonomic Bulletin & Review*, *15*(4), 814–818.

Leibowitz, N., Baum, B., Enden, G., & Karniel, A. (2010). The exponential learning equation as a function of successful trials results in sigmoid performance. *Journal of Mathematical Psychology*, *54*(3), 338–340. https://doi.org/10.1016/j.jmp.2010.01.006

Lemieux, L., Allen, P. J., Franconi, F., Symms, M. R., & Fish, D. K. (1997). Recording of EEG during fMRI experiments: Patient safety. *Magnetic Resonance in Medicine*, *38*(6), 943–952. https://doi.org/10.1002/mrm.1910380614

Leys, C., Klein, O., Dominicy, Y., & Ley, C. (2018). Detecting multivariate outliers: Use a robust variant of the Mahalanobis distance. *Journal of Experimental Social Psychology*, *74*, 150–156. https://doi.org/10.1016/j.jesp.2017.09.011

Liu, Z. (1999). Perceptual learning in motion discrimination that generalizes across motion directions. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(24), 14085–14087. https://doi.org/10.1073/pnas.96.24.14085

Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in Cognitive Sciences*, *16*(10), 511–518. https://doi.org/10.1016/j.tics.2012.08.010

MacCallum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological Methods*, *7*(1), 19–40. https://doi.org/10.1037//1082-989X.7.1.19

Machizawa, M. G., & Driver, J. (2011). Principal component analysis of behavioural individual differences suggests that particular aspects of visual working memory may relate to specific aspects of attention. *Neuropsychologia*, *49*(6), 1518–1526. https://doi.org/10.1016/j.neuropsychologia.2010.11.032

Matzen, L. E., Benz, Z. O., Dixon, K. R., Posey, J., Kroger, J. K., & Speed, A. E. (2010). Recreating Raven's: Software for systematically generating large numbers of Raven-like matrix problems with normed properties. *Behavior Research Methods*, *42*(2), 525–541. https://doi.org/10.3758/BRM.42.2.525

Michel, M. M., & Jacobs, R. A. (2007). Parameter learning but not structure learning: A Bayesian network model of constraints on early perceptual learning. *Journal of Vision*, *7*(1), 4. https://doi.org/10.1167/7.1.4

Miyake, A. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Molenaar, P. C. M. (2004). A Manifesto on Psychology as Idiographic Science: Bringing the Person Back Into Scientific Psychology, This Time Forever. *Measurement: Interdisciplinary Research & Perspective*, *2*(4), 201–218. https://doi.org/10.1207/s15366359mea0204_1

Morey, R. D., Rouder, J. N., & Jamil, T. (2015). *Package 'BayesFactor.'* http://cran.r-project.org/web/packages/BayesFactor/BayesFactor.pdf

Moscatelli, A., Mezzetti, M., & Lacquaniti, F. (2012). Modeling psychophysical data at the population-level: The generalized linear mixed model. *Journal of Vision*, *12*(11), 26–26. https://doi.org/10.1167/12.11.26

Nesbitt, K. T., Farran, D. C., & Fuhs, M. W. (2015). Executive function skills and academic achievement gains in prekindergarten: Contributions of learning-related behaviors. *Developmental Psychology*, *51*(7), 865–878. https://doi.org/10.1037/dev0000021

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–51). Lawrence Erlbaum.

Newell, K. M., Liu, Y. T., & Mayer-Kress, G. (2001). Time scales in motor learning and development. *Psychological Review*, *108*(1), 57–82.

Newell, Karl M., Mayer-Kress, G., Hong, S. L., & Liu, Y.-T. (2009). Adaptation and learning: Characteristic time scales of performance dynamics. *Human Movement Science*, *28*(6), 655–687. https://doi.org/10.1016/j.humov.2009.07.001

Pahor, A., Stavropoulos, T., Jaeggi, S. M., & Seitz, A. (2019). Validation of a matrix reasoning task for mobile devices. *Behavior Research Methods*, *51*(5), 2256–2267. https://doi.org/10.3758/s13428-018-1152-2

Palmer, E. M., Horowitz, T. S., Torralba, A., & Wolfe, J. M. (2011). What are the shapes of response time distributions in visual search? *Journal of Experimental Psychology: Human Perception and Performance*, *37*(1), 58–71. https://doi.org/10.1037/a0020747

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*(0), 96–146. https://doi.org/10.1214/09-SS057

Plant, E. A., & Devine, P. G. (1998). Internal and external motivation to respond without prejudice. *Journal of Personality and Social Psychology*, *75*(3), 811–832. https://doi.org/10.1037/0022-3514.75.3.811

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. *3rd International Workshop on Distributed Statistical Computing*, *124*.

Primi, R., Ferrão, M. E., & Almeida, L. S. (2010). Fluid intelligence as a predictor of learning: A longitudinal multilevel approach applied to math. *Learning and Individual Differences*, *20*(5), 446–451. https://doi.org/10.1016/j.lindif.2010.05.001

Ratcliff, R., & Murdock, B. B. (1976). Retrieval processes in recognition memory. *Psychological Review*, *83*(3), 190–214. https://doi.org/10.1037//0033-295X.83.3.190

Ratcliff, R., & Tuerlinckx, F. (2002). Estimating parameters of the diffusion model: Approaches to dealing with contaminant reaction times and parameter variability. *Psychonomic Bulletin & Review*, *9*(3), 438–481. https://doi.org/10.3758/BF03196302

Raven, J. C. (1998). *Manual for Raven's Progressive Matrices*. Oxford Psychologist Press.

Reddy, P. G., Mattar, M. G., Murphy, A. C., Wymbs, N. F., Grafton, S. T., Satterthwaite, T. D., & Bassett, D. S. (2018). Brain state flexibility accompanies motor-skill acquisition. *NeuroImage*, *171*, 135–147. https://doi.org/10.1016/j.neuroimage.2017.12.093

Ren, X., Schweizer, K., Wang, T., & Xu, F. (2015). The Prediction of Students' Academic Performance With Fluid Intelligence in Giving Special Consideration to the Contribution of Learning. *Advances in Cognitive Psychology*, *11*(3), 97–105. https://doi.org/10.5709/acp-0175-z

Reynolds, J. H., & Heeger, D. J. (2009). The Normalization Model of Attention. *Neuron*, *61*(2), 168–185. https://doi.org/10.1016/j.neuron.2009.01.002

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. https://doi.org/10.1038/323533a0

Seitz, A., & Dinse, H. R. (2007). A common framework for perceptual learning. *Current Opinion in Neurobiology*, *17*(2), 148–153. https://doi.org/10.1016/j.conb.2007.02.004

Seitz, A., Nanez, J. E., Holloway, S., Tsushima, Y., & Watanabe, T. (2006). Two cases requiring external reinforcement in perceptual learning. *Journal of Vision*, *6*(9), 966–973. https://doi.org/10.1167/6.9.9

Shipstead, Z., Lindsey, D. R. B., Marshall, R. L., & Engle, R. W. (2014). The mechanisms of working memory capacity: Primary memory, secondary memory, and attention control. *Journal of Memory and Language*, *72*, 116–141. https://doi.org/10.1016/j.jml.2014.01.004

Simmering, V. R. (2016). Working memory capacity in context: Modeling dynamic processes of behavior, memory, and development. *Monographs of the Society for Research in Child Development*, *81*(3), 1:168.

Simmering, V. R., & Spencer, J. P. (2008). Generality with specificity: The dynamic field theory generalizes across tasks and time scales. *Developmental Science*, *11*(4), 541–555.

Smith, P. L., & Ratcliff, R. (2004). Psychology and neurobiology of simple decisions. *TRENDS in Neurosciences*, *24*(3), 161–168.

Snoddy, G. S. (1926). Learning and stability: A psychophysiological analysis of a case of motor learning with clinical applications. *Journal of Applied Psychology*, *10*(1), 1–36. https://doi.org/10.1037/h0075814

Stanovich, K. E. (1986). Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy. *Reading Research Quarterly*, *21*(4), 360–407. https://doi.org/10.1598/RRQ.21.4.1

Stratton, S. M., Liu, Y.-T., Hong, S. L., Mayer-Kress, G., & Newell, K. M. (2007). Snoddy (1926) revisited: Time scales of motor learning. *Journal of Motor Behavior*, *39*(6), 503–515. https://doi.org/10.3200/JMBR.39.6.503-516

Thelen, E., & Smith, L. B. (1994). *A Dynamic Systems Approach to the Development of Cognition and Action*. MIT Press.

Vaci, N., Edelsbrunner, P., Stern, E., Neubauer, A., Bilalić, M., & Grabner, R. H. (2019). The joint influence of intelligence and practice on skill development throughout the life span. *Proceedings of the National Academy of Sciences*, 201819086. https://doi.org/10.1073/pnas.1819086116

Voelke, A. E., Troche, S. J., Rammsayer, T. H., Wagner, F. L., & Roebers, C. M. (2014). Relations among fluid intelligence, sensory discrimination and working memory in middle to late childhood – A latent variable approach. *Cognitive Development*, *32*, 58–73. https://doi.org/10.1016/j.cogdev.2014.08.002

Wagenmakers, E.-J., & Farrell, S. (2004). AIC model selection using Akaike weights. *Psychonomic Bulletin & Review*, *11*(1), 192–196. https://doi.org/10.3758/BF03206482

Wang, T., Ren, X., & Schweizer, K. (2017). Learning and retrieval processes predict fluid intelligence over and above working memory. *Intelligence*, *61*, 29–36. https://doi.org/10.1016/j.intell.2016.12.005

Wang, X., Zhou, Y., & Liu, Z. (2013). Transfer in motion perceptual learning depends on the difficulty of the training task. *Journal of Vision*, *13*(7), 5–5. https://doi.org/10.1167/13.7.5

Watanabe, T., Náñez, J. E., & Sasaki, Y. (2001). Perceptual learning without perception. *Nature*, *413*(6858), 844–848. https://doi.org/10.1038/35101601

Watanabe, T., & Sasaki, Y. (2015). Perceptual Learning: Toward a Comprehensive Theory. *Annual Review of Psychology*, *66*(1), 197–221. https://doi.org/10.1146/annurev-psych-010814-015214

Westermann, G., Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M. W., & Thomas, M. S. C. (2007). Neuroconstructivism. *Developmental Science*, *10*(1), 75–83. https://doi.org/10.1111/j.1467-7687.2007.00567.x

Westlye, L. T., Grydeland, H., Walhovd, K. B., & Fjell, A. M. (2011). Associations between Regional Cortical Thickness and Attentional Networks as Measured by the Attention Network Test. *Cerebral Cortex*, *21*(2), 345–356. https://doi.org/10.1093/cercor/bhq101

Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E.-J. (2011). Statistical Evidence in Experimental Psychology: An Empirical Comparison Using 855 *t* Tests. *Perspectives on Psychological Science*, *6*(3), 291–298. https://doi.org/10.1177/1745691611406923

Wright, T. P. (1936). Factors Affecting the Cost of Airplanes. *Journal of the Aeronautical Sciences*, *3*(4), 122–128. https://doi.org/10.2514/8.155

Xu, K., Nosek, B., & Greenwald, A. G. (2014). Data from the Race Implicit Association Test on the Project Implicit Demo Website. *Journal of Open Psychology Data*, *2*(1), e3. https://doi.org/10.5334/jopd.ac

Appendices

Appendix 1. Simulations of learning within a memory task

Data generation function (R):

```
function (pars = list(g1 = c(p_s = 0.6, p_r = 4, p_a = 0.75),
    g2 = c(p_s = 0.6, p_r = 6, p_a = 0.85)), nTrials = 100, chanceLevel = 0.5,
    nSims = 10, binomN = 1)
{
    simDat <- data.frame()
    trialNums <- (1:nTrials)
    for (. in 1:nSims) {
        for (curGroup in names(pars)) {
            theta <- pars[[curGroup]]["p_a"] + (pars[[curGroup]]["p_s"] -
                pars[[curGroup]]["p_a"]) * exp((1 - trialNums)/(2^pars[[curGroup]]["p_r"]))
            binResp <- rbinom(length(theta), binomN, theta)/binomN
            simDat <- rbind(simDat, data.frame(subID = paste0(sample(letters,
                7, replace = T), collapse = ""), group = curGroup,
                p_s = pars[[curGroup]]["p_s"], p_r = pars[[curGroup]]["p_r"],
                p_a = pars[[curGroup]]["p_a"], trialNum = trialNums,
                theta = theta, resp = binResp, sqEr = (binResp -
                  theta)^2, binomN = binomN))
        }
    }
    return(simDat)
}
```

Model formula (brms package)

```
resp ~ inv_logit(thAsym) + (inv_logit(thStart) - inv_logit(thAsym)) * exp((1 -
trialNum)/(2^thRate))
        thAsym ~ group + (1 | subID)
        thStart ~ group + (1 | subID)
        thRate ~ group + (1 | subID)
        phi ~ (1 | subID)
```

Appendix 2. Kattner, Cochrane & Green (2017)

Trial-dependent psychometric functions accounting for perceptual learning in 2-AFC discrimination tasks

Florian Kattner
Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA
Institute of Psychology, Technische Universität Darmstadt, Darmstadt, Germany

Aaron Cochrane
Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

C. Shawn Green
Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA

The majority of theoretical models of learning consider learning to be a continuous function of experience. However, most perceptual learning studies use thresholds estimated by fitting psychometric functions to independent blocks, sometimes then fitting a parametric function to these block-wise estimated thresholds. Critically, such approaches tend to violate the basic principle that learning is continuous through time (e.g., by aggregating trials into large "blocks" for analysis that each assume stationarity, then fitting learning functions to these aggregated blocks). To address this discrepancy between base theory and analysis practice, here we instead propose fitting a parametric function to thresholds from each individual trial. In particular, we implemented a dynamic psychometric function whose parameters were allowed to change continuously with each trial, thus parameterizing nonstationarity. We fit the resulting continuous time parametric model to data from two different perceptual learning tasks. In nearly every case, the quality of the fits derived from the continuous time parametric model outperformed the fits derived from a nonparametric approach wherein separate psychometric functions were fit to blocks of trials. Because such a continuous trial-dependent model of perceptual learning also offers a number of additional advantages (e.g., the ability to extrapolate beyond the observed data; the ability to estimate performance on individual critical trials), we suggest that this technique would be a useful addition to each psychophysicist's analysis toolkit.

Introduction

One common assumption, instantiated in numerous theoretical models in the domains of psychology, neuroscience, and computer science, is that learning is a continuous function of experience. For example, this assumption underlies all models that use some form of a delta rule procedure (Casey & Sowden, 2012; Rumelhart, Hinton, & Williams, 1986; Spratling & Johnson, 2006). Here, in each learning epoch, the learner makes a prediction regarding the correct output, and then receives feedback as to the true correct output. The learner then computes the difference between their prediction and the true correct output and uses this to update the next prediction. When done repeatedly over time, this process will tend to gradually move the learner's predictions into alignment with the true correct outputs. Learning is also modeled as a continuous process in many purely associative learning models (Bejjanki, Beck, Lu, & Pouget, 2011; Guenther, Ghosh, & Tourville, 2006; Rosenblatt, 1958; Spratling & Johnson, 2001). These models regularly use some form of Hebbian learning principle, wherein the strength of the connection between two nodes is updated after each learning epoch by an amount proportional to the extent to which the two nodes were simultaneously active during the learning epoch. Finally, Bayesian learning models are inherently continuous in nature, as each observed training example increases or decreases the probability that a particular estimate/hypothesis is correct (by an amount that depends on the strength of the evidence

Citation: Kattner, F., Cochrane, A., & Green, C. S. (2017). Trial-dependent psychometric functions accounting for perceptual learning in 2-AFC discrimination tasks. *Journal of Vision*, *17*(11):3, 1–16, doi:10.1167/17.11.3.

doi: 10.1167/17.11.3 Received April 4, 2016; published September 6, 2017 ISSN 1534-7362

provided in the training example and the prior probability that the particular estimate/hypothesis was correct; (Jacobs & Kruschke, 2011; Michel & Jacobs, 2007). Thus, although the general spirit as well as the detail-level implementation of these models may vary substantially, all instantiate this same basic principle that learning is a continuous process wherein small changes in ability accumulate via experience (Lu, Hua, Huang, Zhou, & Dosher, 2011; Mazur & Hastie, 1978; Petrov, Dosher, & Lu, 2005).

Not surprisingly, models in the field of perceptual learning share this same fundamental assumption that learning mechanisms are inherently incremental (Herzog & Fahle, 1998; Law & Gold, 2009; Lu et al., 2011; Petrov et al., 2005; Poggio, Fahle, & Edelman, 1992; Sotiropoulos, Seitz, & Seris, 2011; Vaina, Sundareswaran, & Harris, 1995; Zhaoping, Herzog, & Dayan, 2003). Indeed, theoretical models in this domain frequently have at their core one of the three broad types of learning rules/processes above (i.e., delta rule, associative/Hebbian, Bayesian). Interestingly, perceptual learning is even posited to be continuous in conditions that may not initially seem to easily support such learning. Take, for instance, the case of block feedback (Herzog & Fahle, 1997). In training tasks that employ block feedback, participants do not receive feedback regarding their accuracy after each trial. Instead, they are given their average accuracy across the entire previous block of trials after the block is finished. This poses a challenge for many of the models aforementioned, which require an explicit error signal to update behavior (i.e., the type of signal that would typically come from trial-by-trial feedback). However, one influential theoretical model in the field of perceptual learning produces continuous changes in performance even in this block feedback case. In this model, participants can use external and internal signals to update behavior. If, as is true in block feedback designs, there is not an external learning signal available to drive learning on each trial, the model will instead use internal estimates to alter performance continuously (with the internal signals being updated whenever external feedback is provided; Liu, Dosher, & Lu, 2014).

Thus, given the fact that essentially all theoretical models in the domain of perceptual learning suggest that learning should be continuous with experience, it is interesting to note that, in most behavioral experiments in this domain, learning is accounted for in a discontinuous manner. Rather than modeling changes in behavior as a continuous process using completely trial-dependent parameters, learning is often accounted for by first computing performance in discrete "blocks" of trials and then using the differences across those blocks (or fitting a parametric function to block performance as the measure of learning; Ball & Sekuler, 1987; Beard, Levi, & Reich, 1995; Dosher & Lu, 1998; Fahle & Edelman, 1993; Fahle & Morgan, 1996; Fendick & Westheimer, 1983; Gantz, Patel, Chung, & Harwerth, 2007; Seitz, Nanez, Holloway, Tsushima, & Watanabe, 2006; Yu, Klein, & Levi, 2004). For example, in one common method, the learning data is first subdivided into discrete blocks of trials, with the block size typically being based upon the experimental methods that were employed and ranging anywhere from 50–700 trials; (Ball & Sekuler, 1987; Fahle & Morgan, 1996). Then, a psychometric function is fit to the data within each block (e.g., logistic, Weibull, or cumulative Gaussian; Coates & Chung, 2014; Crist, Kapadia, Westheimer, & Gilbert, 1997; Fahle & Edelman, 1993) and a threshold value is calculated (e.g., 79% threshold). The difference in this threshold value from early blocks in training to late blocks in training is then used as the quantification of learning. In another common method, the threshold values for blocks are parametrically fit with a monotonically decreasing function (e.g., power, exponential; Astle, Blighe, Webb, & McGraw, 2015; Chung, 2011; Coates & Chung, 2014; Herzog & Fahle, 1997, 1999; Levi, Polat, & Hu, 1997; Matthews, Liu, Geesaman, & Qian, 1999; for a review, see Dosher & Lu, 2007).

Critically though, one important implicit assumption of such fitting procedures is that the parameters of the function are ***not*** changing over the block of trials being considered (e.g., the fitting in these cases necessarily assume that the data is generated by a constant level of performance). Thus, even when using parametric fits to block thresholds, performance is assumed to be stationary within each block and the most precise estimate of performance and learning is at the aggregated block level (alternatively, each block threshold must be taken to represent a particular trial in the block, e.g., the middle trial or the first trial). The process of fitting a learning function to block thresholds is problematic itself due to the errors inherent in sequentially modeling hierarchical data; see, e.g., Moscatelli, Mezzetti, & Lacquaniti, 2012.

This same implicit assumption regarding within-block stationarity of performance also underlies essentially all adaptive techniques for quickly estimating thresholds (e.g., staircases, PEST, QUEST, etc.; see Treutwein, 1995). Indeed, using an adaptive technique to estimate a threshold makes little sense if the threshold is actively changing during the estimation. Finally, this assumption of stationarity is present in any statistics that simply aggregate performance over an entire block without fitting. This includes analyses built upon signal detection theory (e.g., *d'* analysis assumes that the particular pattern of hits and false alarms across a block of trials is driven by a constant sensitivity) as well as any technique wherein performance is quantified as a simple average over blocks of

trials (e.g., percent correct). In essence, by using the aforementioned approaches, participants are being modeled as not changing at all *within* blocks of trials and instead are only free to improve *in between* blocks of trials (i.e., learning in a stepwise fashion). Because such a stepwise function is in direct contrast to our theoretical understanding of learning as a continuous function characterizing the relation between improvement and experience, in the present paper we present a new method of analyzing perceptual learning data to account for continuous changes in performance as a function of experience. Specifically, we employ a standard psychometric function whose parameters are allowed to change continuously through time. This is conceptually identical to fitting a psychometric function to all data points as a single block, but parameterizing nonstationarity rather than assuming within-block stationarity (Fründ, Haenel, & Wichmann, 2011). By fitting the psychometric function to the largest possible "block" (i.e., all trials) we reduce noise introduced by factors other than perceptual ability, and by estimating learning as a function of the smallest possible "block" (i.e., each individual trial) our estimates better reflect the continuous nature of learning. In addition, we improve upon parametric fits to block estimates by requiring fewer free parameters, while also providing simultaneous fits to stimulus (threshold) and time (learning) dimensions.

Here, we show that our continuous time-parametric model provides a better fit, without overfitting, to perceptual learning data than the more traditional trial-independent, nonparametric approach of fitting psychometric functions to data in consecutive blocks of trials. This is perhaps not surprising given that block models necessarily take a functional form that is inconsistent with our beliefs about actual human learning. Furthermore, in addition to simply providing a better fit to perceptual learning data, the continuous time-parametric model also offers a number of other empirical (e.g., more accurate extrapolation of performance) and theoretical advantages (e.g., ability to use all data in assessing the functional form of learning; provides a natural method for estimating certain important trials, such as the first and last trial of training) over standard nonparametric block fitting or parametric fits to blocks. We therefore suggest it will be a valuable addition to every psychophysicist's toolkit.

Method

Perceptual learning data/tasks

Data from two different standard perceptual learning tasks was used in the current analysis. Both data sets, including one examining orientation discrimination training ($N = 7$) and one examining stereoacuity training ($N = 7$), overlap with previously published data sets (Green, Kattner, Siegel, Kersten, & Schrater, 2015; Snell, Kattner, Rokers, & Green, 2015). In the following material, we briefly describe the basic training methods for the data that is considered (Note: For each of the following tasks, participants underwent brief pretests without feedback prior to training on both the to-be-trained task as well as various transfer measures; however, because the focus of the current manuscript is on fitting learning curves this data is not considered).

Orientation discrimination training methods

For full task methods, see Green et al. (2015). Briefly, in the orientation discrimination training task, on each trial, participants were presented with a central "T" (either upright or upside down) as well as a full-contrast Gabor patch at an eccentricity of 10° below the "T." The orientation of the Gabor was drawn from a uniform random distribution between 30° and 60°. After the stimuli were presented, the participants were required to first respond to the orientation of the "T" (by pressing the "w" or "s" key for upright or upside down, respectively), and then to the orientation offset of the Gabor relative to 45° (by pressing the right or left arrow key for clockwise or counterclockwise, respectively). Participants completed 3,800 such trials, distributed over four different days.

Stereoacuity training methods

For full task methods, see Snell et al. (2015). In the stereoacuity training task, on each trial, two white three-dimensional rectangles, offset relative to one another in depth, were presented. The size of the offset was drawn from a uniform distribution between 0 and 60 arcsec. The participants' task was to indicate which square appeared closer in depth. Participants completed 7,500 of such trials, distributed across five different days.

Parametric model of continuous perceptual learning

Because both tasks involved participants making two-alternative forced choice (2-AFC) decisions on stimuli that varied in signal intensity, the continuous model was built upon a generalized psychometric function (Equation 1), relating an observer's responses to stimulus intensity x (e.g., orientation or stereo offsets; see Supplemental Materials for additional fitting details related to software, etc.):

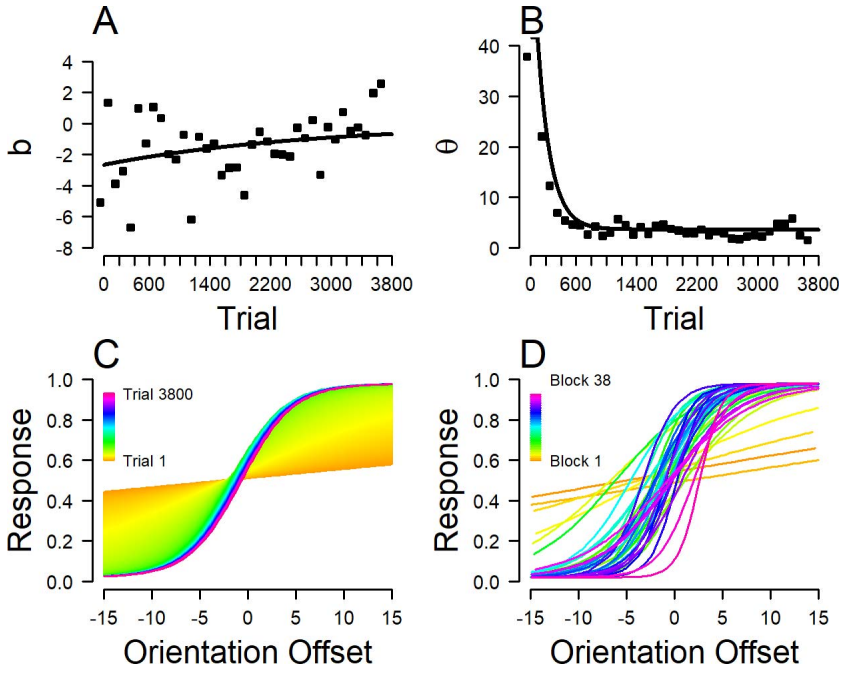


Figure 1. Illustration of the parameters *b* and θ as fit using the block and continuous models of psychophysical data with an exemplar orientation discrimination subject. (A) Change of the bias value *b* fit within 38 independent successive blocks (black squares) or as a continuous function of trial number (solid line). (B) Change of the threshold value θ fit as 38 independent successive blocks (black squares) and as a continuous function of trial number (solid line). (C) The resulting trial-dependent psychometric function as estimated with the continuous model (0 = counterclockwise, 1 = clockwise). As is clear, the continuous approach models performance as changing smoothly through time. (D) The resulting psychometric functions in 38 independent 100-trial blocks of training (0 = counterclockwise, 1 = clockwise). This approach is, in essence, only allowing for changes in performance across blocks of trials.

$$f(x; b, \theta) = \lambda + \frac{1 - 2\lambda}{1 + k^{\frac{b-x}{\theta}}} \quad (1)$$

Equation 1 contains two constants: The parameter k was defined as 3.76 (= 0.79/0.21) in order to estimate 79% thresholds. λ refers to a lapse parameter. In essence, this can be considered the probability that a participant will show a total "lapse" of attention in which case his/her performance will be unrelated to the stimulus that was presented; in practice, this is used to account for trials at very high levels of stimulus strength that the participant nonetheless answers incorrectly (see Klein, 2001). This lapse value was held constant at 0.02. (Note that values for the lapse parameters between 0 and 0.1 were assessed—for the most part, the particular value did not affect the quality of fits, although there was a tendency for the largest values to somewhat degrade the fits.) The two remaining parameters b and θ refer to the bias and threshold of the psychometric function, respectively. To account for continuous perceptual learning, these two parameters (b and θ) were themselves fit as functions of time (t; see Equations 2 and 3 as follows). Because the focus of this article is not on the exact functional form of those parameters in relation to time (we note that identifying the exact functional form of the change function is beyond the scope of this paper; see Discussion and Supplemental Materials, and Kattner, Cochrane, Cox, Gorman, & Green, 2017 for an alternative parameterization), we chose to model bias as a two-parameter exponential function of time (Equation 2) and threshold as a three-parameter exponential function of time (Equation 3).

$$b(t) = b_0 \cdot e^{-\frac{t}{b_1}} \quad (2)$$

$$\theta(t) = pre - (pre - post)e^{-\frac{t}{\tau}} \quad (3)$$

Continuous perceptual learning in the two sets of data can thus be accounted for by a psychometric function with two constants (k and λ) and five independent parameters, with b_0 referring to initial bias, *pre* to the initial threshold, *post* to the final asymptote of the threshold, and two slope parameters for bias and threshold (b_1 and τ, respectively).

The relationship between trial-dependent parameter estimates (bias and threshold) and the time-evolving psychometric function is illustrated in Figure 1 (panels A, B, C) for an exemplar participant trained on the orientation discrimination task.

Nonparametric (block) model of perceptual learning

As a standard against which the continuous time parametric model could be compared, we also fit the data via a block-based method commonly used (Dosher & Lu, 2000) in the perceptual learning field (see Figure 1, panel D). For each participant and task, the data was first divided into blocks of 100 trials each. A single logistic function (Equation 1; two free parameters) was then fit to each block. The relationship between block-by-block parameter estimates (bias and threshold) and the resulting discrete psychometric functions in each block are illustrated in Figure 1 (panels A, B, and D) for an exemplar orientation discrimination participant.

Hybrid model (see Supplemental Materials)

Our main interest in this paper is in comparing the two analysis approaches described already (approach #1: the nonparametric block model where thresholds are fit to blocks of aggregated data thus not assuming a trial-dependent change in parameters, e.g., Crist et al., 1997; Fahle & Edelman, 1993; approach #2: the continuous time parametric model where thresholds are fit by considering trial-by-trial changes in the parameters of the psychometric function). However, it is worth noting that there is a third approach that is, to some extent, an intermediate between a fully time-continuous model and a fully block model. Namely, it is possible to account for perceptual learning by first aggregating data within discrete blocks of trials and then fitting a continuous model to these "block-averaged" response probabilities. Because parametric fitting to block thresholds is a common approach to analyzing data in the literature (e.g., Chung, 2011; Coates & Chung, 2014; Fründ et al., 2011; see Dosher & Lu, 2007 for a review of earlier studies) we present the results of this type of "hybrid" model in the Supplemental Materials along with several other alternative models , all of which use the same learning functions (Equations 2 and 3; see Discussion).

Comparing continuous parametric and block nonparametric analyses

In comparing the continuous parametric and block-based nonparametric analysis approaches, we examined several basic aspects of model quality. The first was simply to examine how well the model captures the full pattern of data. To this end, after fitting both the continuous time and the block model to the full training data for each participant, several measures were assessed, including Akaike and Bayesian information criteria (AIC and BIC, respectively). Both metrics provide estimates of the quality of a model relative to other models. More specifically, both involve a calculation of the likelihood function (i.e., probability of the data given the model; see Equation 4, with r being the observed binary responses) that is then penalized based upon the number of parameters in the model (penalty term $= [-2\log L + kp]$, with L being the likelihood function, p being the number of parameters in the model, and k being an additional penalty that differs for AIC and BIC). In particular, with large numbers of parameters, this penalty is greater in the case of BIC than AIC (the term k is set to 2 for AIC and $\log(p)$ for BIC). In addition to AIC and BIC, we also calculated χ^2 measures (accounting for the discrepancy between theoretical/fitted and observed data, see Supplementary Materials for equation; cf. Klein, 2001).

$$logL = (\Sigma \log(f(x; b, \theta))r + (1 - \log(f(x; b, \theta)))(1 - r)) \quad (4)$$

Second, one major concern in essentially all data modeling is related to overfitting—when the model captures random fluctuations/noise in the data rather than only capturing true signal (i.e., in the case of perceptual learning, the true signal would be actual changes in performance). Overfitting becomes a greater and greater concern as models increase in complexity (e.g., increases in the number of free parameters; see Figure 2). To account for overfitting, the quality of each model fit was assessed in a standard train/test procedure. Specifically, the models were first fit to data on the odd trials only (i.e., 1,900 and 3,750 orientation and stereo discrimination trials, respectively). The quality of the resulting model fit was then assessed with respect to data on the even trials by calculating AIC, BIC, and χ^2. The rationale here is that the training set (i.e., the odd trials) should be generated from the same basic perceptual sensitivity as the test set (i.e., the even trials). Thus, by examining how well the fits derived from the training set (odd trials) match the data in the test set (even trials), we can assess the extent to which the two competing models properly fit the data, without overfitting the data.

Third, 95% confidence intervals around the fits were computed via bootstrapping (i.e., drawing random samples with replacement from the individual trial sequences, fitting psychometric functions and calculating threshold of each sample, and determining the confidence intervals from the 97.5% and 2.5% percentiles; see Supplemental Materials for additional detail on the bootstrapping procedure). The width of the confidence intervals was then contrasted between the models. Like already mentioned, this provides an estimate of model quality, again, in particular the

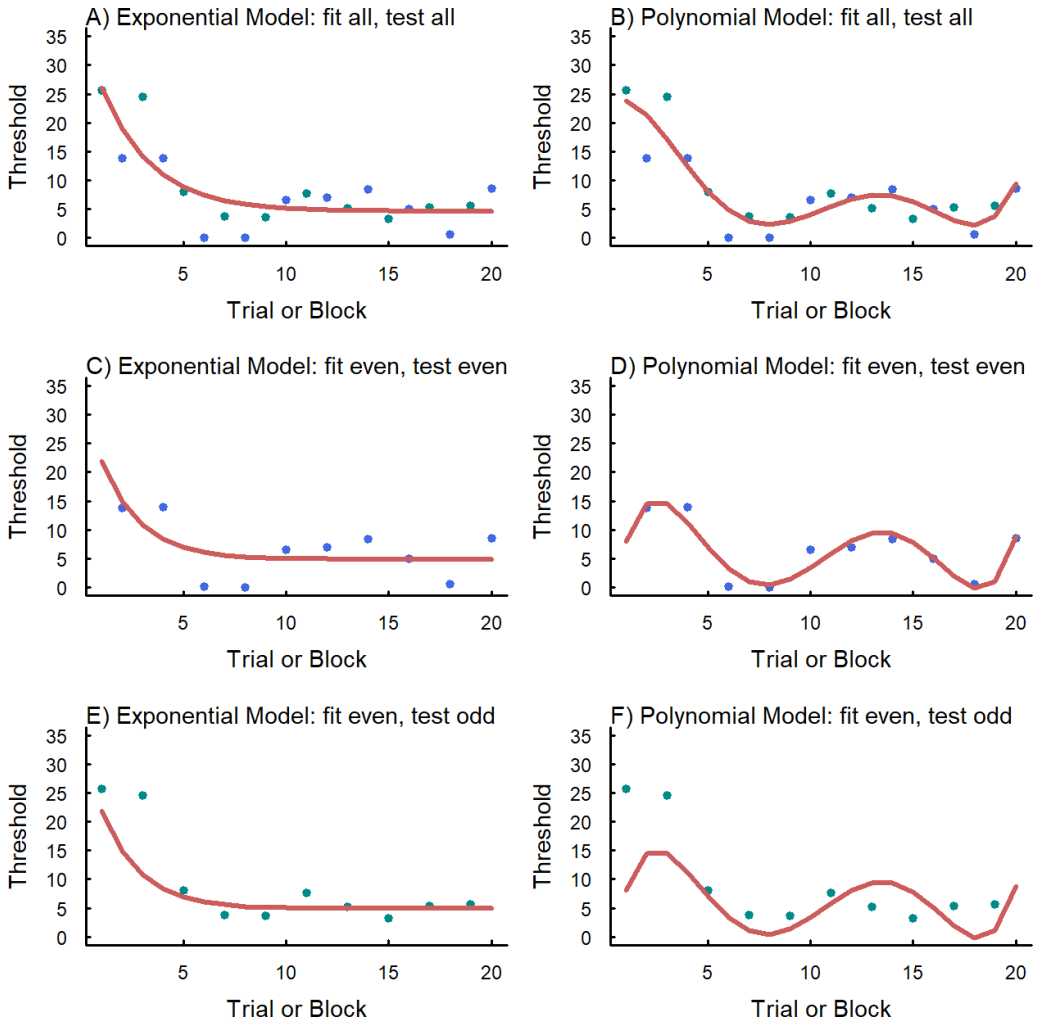


Figure 2. Illustration of how overfitting can be detected. (A) Fit to an arbitrary time series data using a less complex model (exponential). (B) Fit to the same arbitrary time series data as in A, but using a more complex model (high-order polynomial). In examining how well the two models fit the data in A and B, it is clear that both do a reasonable job of predicting the position of the data points with the more complex model, if anything, doing a better job. (C) Fit to just the even trials of the time series data using the less complex model. (D) Fit to just the even trials of the time series data using the more complex model. Again, both models do a reasonable of predicting the position of the data points. (E) The fit on the even trials from the less complex model continues to do a good job of predicting the position of the untrained (odd) trials. (F) The fit on the even trials from the more complex model does a quite poor job of predicting the position of the untrained (odd) trials. This is the hallmark of an overfitting model. It does a good job of making predictions about the trained data set, but performance is markedly poorer for the untrained data.

extent to which the model is susceptible to overfitting. A model that is more prone to overfitting might produce markedly different estimates depending on the particular (perhaps idiosyncratic) set of trials that is considered. A model that is less prone to overfitting should produce essentially the same estimates regardless of the exact set of trials that is considered.

Fourth, both the models were fit only to an initial portion of the data and then used to predict fits of the remaining trials. This tests a final critical aspect of perceptual learning data analyses – the ability to make a forward prediction. Specifically, thresholds were fit to the responses on the initial 1,500 orientation and 3,000 stereo discrimination trials, respectively, and the fitted thresholds were then extrapolated over the remaining trials of the respective task. The extrapolated estimates were then contrasted against the true data by calculating the AIC, BIC, and χ^2 measures. Note that the nonparametric block approach provides no natural method of extrapolation (as it does not implement any particular functional form). Thus, for the block model, the extrapolation was obtained by fitting continuous functions (Equations 2 and 3) to the threshold and bias parameter estimates obtained for each block (i.e., with constant biases and thresholds for all trials within a block) of the initial subset of the data. These parameter functions can then be used to predict threshold for the remaining blocks.

	Continuous model				Block model			
Subject	*logL*	*AIC*	*BIC*	χ^2	*logL*	*AIC*	*BIC*	χ^2
O1	−1477	2964	2995*	3998*	−1403*	2957*	3155	4094
O2	−1643	3296	3327*	3723	−1508*	3169*	3367	3388*
O3	−1626	3263	3294*	3829	−1540*	3232*	3430	3652*
O4	−1086	2182*	2213*	3415	−1029*	2210	2408	3147*
O5	−2380	4770	4801*	3810*	−2268*	4687*	4885	3898
O6	−1281	2571	2603*	4335*	−1198*	2549*	2747	4413
O7	−1671	3351	3383*	3747*	−1532*	3215*	3413	4102
S1	−3019	6049	6083*	7929	−2838*	5975*	6366	7430*
S2	−3972	7955	7989*	7638	−3825*	7949*	8340	7534*
S3	−2357	4724	4758*	9201	−2157*	4613*	5004	7881*
S4	−4271	8551*	8586*	7713*	−4141*	8582	8973	7766
S5	−2808	5626	5661*	7815	−2650*	5600*	5991	7780*
S6	−4369	8748*	8783*	7564	−4231*	8762	9153	7562*
S7	−3590	7191	7225*	7732	−3423*	7146*	7537	7499*

Table 1. Overall analysis of model fits: Goodness of fit metrics (log likelihoods, AIC, BIC, and χ^2) for the continuous model (np = 5) and block model (*np* = 76 and *np* = 150, respectively) fit to *nd* = 3,800 orientation discrimination trials (subjects O1–O7) and *nd* = 7,500 stereo discrimination trials (subjects S1–S7). Asterisks indicate the model with the best relative fit to the data (note that this does not indicate "statistical significance" in the traditional null-hypothesis sense).

Results

Overall fit

The models were first fit to responses in all 3,800 and 7,500 trials of orientation and stereo discrimination training, respectively. The resulting goodness of fit metrics are summarized in Table 1 (see the Supplemental Materials for information on the hybrid models). Both models clearly fit the data well overall. Of interest is that the block model tends to show better performance as determined by AIC values (better for 11 out of 14 participants), whereas the continuous model shows better performance as determined by BIC values (better for 14 out of 14 participants). Less consistent support for either model has been obtained with the χ^2 metric (i.e., the discrepancy between observed and predicted data was smaller for the continuous model in five out of 14 participants). Given that the major difference between AIC and BIC is the extent to which greater model complexity is penalized, it was next of interest to examine whether the better performance seen in AIC values in the block model is due to overfitting.

Test for overfitting: Train/test analysis

For both tasks, the goodness of fit of the models was evaluated by fitting the models to the responses on odd trials and testing with regard to the responses on even trials. Table 2 shows the resulting AIC, BIC, and χ^2 metrics for the block and continuous models fit to the data of each participant (see Supplemental Materials for the hybrid models). As can be seen in Table 2, the test data was consistently fit better by the continuous model (providing a markedly better fit for all 14 participants regardless of the measure of model fit employed). It is remarkable, given that the continuous model only has a fraction of the number of free parameters that the block model has, that this more-parsimonious fitting method is clearly a better fit to the test data. This pattern of results suggests very strongly that while the nonparametric block approach appeared to do a reasonable job when examining the fit to the full data set, it was very likely in fact dramatically overfitting the data (i.e., block threshold estimates are likely fitting some noise). Meanwhile, the continuous model, which is (a) far less complex in terms of number of parameters than the block model and (b) instantiates a strong belief about the manner in which the data should be generated, appears to be considerably less susceptible to overfitting. Just as larger blocks will reflect perceptual ability more accurately by averaging over more noise in the data, fitting all of the data as one nonstationary block (i.e., the continuous parametric fit) minimizes the influences of noisy data on threshold estimates.

Learning gains and confidence intervals

The amount of perceptual learning in both discrimination tasks can be quantified (with both models) by subtracting the final threshold estimates from the initial threshold estimate (e.g., first 100 trials minus

	Continuous model				Block model			
Subject	*logL*	*AIC*	*BIC*	χ^2	*logL*	*AIC*	*BIC*	χ^2
O1	−792*	1594*	1621*	2201*	−882	1916	2061	4459
O2	−820	1651*	1678*	1852*	−814*	1780	1926	2481
O3	−785*	1580*	1607*	1802*	−807	1766	1911	2313
O4	−520*	1050*	1078*	1555*	−564	1279	1425	2405
O5	−1189	2387*	2415*	1900*	−1181*	2514	2660	2345
O6	−628*	1267*	1294*	2175*	−672	1496	1642	3341
O7	−841*	1692*	1720*	1892*	−866	1885	2030	3328
S1	−1496*	3002*	3033*	3858*	−1578	3455	3742	5923
S2	−2008*	4026*	4057*	3855*	−2084	4468	4754	5105
S3	−1205*	2419*	2450*	4819*	−1243	2786	3073	5995
S4	−2157*	4324*	4355*	3869*	−2226	4752	5039	4939
S5	−1365*	2739*	2770*	3753*	−1413	3127	3413	4669
S6	−2184*	4379*	4410*	3797*	−2264	4827	5114	4738
S7	−1788*	3586*	3617*	3788*	−1836	3972	4259	5042

Table 2. Overfitting analysis for the orientation (subjects O1–O7) and stereo (subjects S1–S7) discrimination data. The continuous ($np = 5$) and the block model ($np = 76$ and $np = 150$) were fit to $nd = 1{,}900$ and $nd = 3{,}750$ even orientation and stereo discrimination trials, respectively. Models were then tested with regard to the same number of odd trials. Asterisks indicate the model with the best relative fit to the data.

last 100 trials; note that the initial orientation and stereo discrimination threshold estimates were constrained to a maximum of 90° and 1.5 arcmin, respectively). The overall amount of learning estimated with the two models was almost identical: For the orientation discrimination task, the block model found threshold improvements of $M = 19.34°$ ($SD = 23.67$), whereas the continuous model estimates an average reduction in threshold of $M = 19.28°$ ($SD = 20.94$). For the stereo discrimination task, improvements of 25.07 arcsec ($SD = 29.49$) were found with the block model, whereas the continuous model suggests a threshold decrease of 25.88 arcsec ($SD = 22.16$). The estimated improvements did not differ significantly between models for either task, $p = 0.71$ and $p = 0.90$, respectively (using nonparametric Wilcoxon rank tests; similar nonsignificant results are found using t tests).

However, differences were found with regard to the confidence of the fitted thresholds that were reached as a result of perceptual learning. The individual 79% thresholds and 95% confidence intervals, as obtained with the continuous and with the block model, are illustrated in Figures 3 and 4 for the orientation and stereo discrimination task, respectively. For the orientation discrimination data, a one-sample Wilcoxon signed-rank test revealed that the confidence intervals orientation discrimination thresholds estimated for the final 100-trial block were significantly smaller with the continuous model ($M = 1.49°$; $SD = 1.27°$) than with the block model ($M = 4.59°$; $SD = 3.27°$), $p = 0.03$ (see Figure 3). Likewise, the confidence intervals of the estimated stereo discrimination thresholds in the last block were significantly lower with the continuous model ($M = 7.51$ arcsec; $SD = 7.96$ arcsec) than with the block model ($M = 25.93$ arcsec; $SD = 20.09$ arcsec), $p = 0.02$ (see Figure 4).

For both the continuous and the block model, the average CI in the last block was subtracted from the average CI in the first block for each data set in order to quantify how well both models capture learning-related decreases in uncertainty of the two models. With the continuous model, the median CI decrement was 71.80° for the orientation discrimination task, and 17.11 arcsec for the stereo discrimination data. In contrast, with the block model, the CI decrements were 15.56° and 11.67 arcsec for the orientation and stereo discrimination tasks, respectively.

Extrapolation analysis

The qualities of the continuous and the block models were further evaluated by fitting both models only to an initial portion of the training data (i.e., 1,500 and 3,000 trials for the orientation and stereo discrimination tasks, respectively), and then extrapolating the thresholds for the remaining trials of each training task, based on the fitted models. For the majority of data from both psychophysical tasks, the thresholds extrapolated based on the continuous model fit the actual data better than did the thresholds extrapolated based on the block model. The exact goodness-of-fit measures for the block and continuous models extrapolated to data from the two tasks are summarized in Table 3 (see the Supplemental Materials for the hybrid model). As

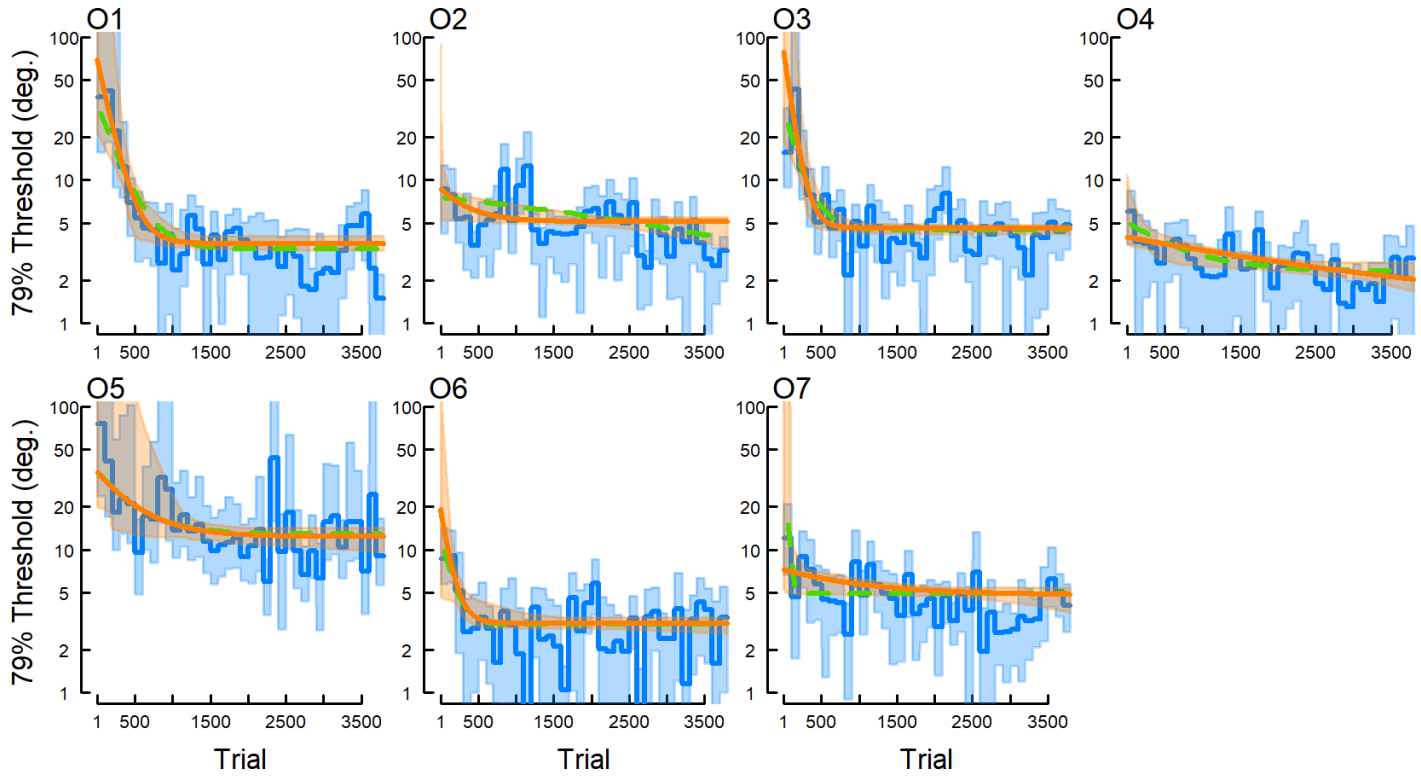


Figure 3. Individual orientation discrimination thresholds based on the 38 separate logistic fits (Equation 1) to 100-trial blocks (block model; blue lines), and the five-parameter continuous model (orange lines). The shaded areas represent the respective bootstrapped 95% confidence intervals for the block and continuous fits, respectively. The dashed green line refers to the hybrid model thresholds (see Supplemental Materials).

expected, for six out of seven orientation discrimination participants and for all stereo discrimination participants, the BIC metrics of extrapolated psychometric functions reached a better fit with the continuous model than with the block model. Only for one participant (O6), responses on the first 1,500 trials could be extrapolated better with the block model (based on BIC), probably due to a slower rate of learning (i.e., the continuous model may not have identified a consistent decrease in threshold during this period; see Figure 3).

Discussion

The majority of theoretical models across many domains of psychology, including the domain of perceptual learning, consider learning to be a process that occurs continuously with experience. However, despite this, most empirical studies in this domain have modeled learning as arising via a discontinuous process. Indeed, participant data in this field is nearly always first separated into distinct blocks of trials for analysis, with block sizes typically being guided by experimental decisions (e.g., how many trials are feasible per day). From there, whether the analyses involve data fitting (e.g., fitting performance across each block with a psychometric function) or simple aggregation/computation (e.g., percent correct across the block; *d'* across the block), all share the implicit assumption that there is no significant change in performance *within* blocks and instead only allow for learning to occur *in-between* blocks.

Given the clear mismatch between the theoretical and analytical approaches in this domain, here we sought to develop a method to bring these approaches into better alignment. Specifically, using data collected from two perceptual learning experiments, we compared two main analytical data fitting methods—the standard nonparametric approach (fitting psychometric functions to blocks of trials) and a new continuous time parametric approach that allows for a trial-dependent continuous change in the parameters of the psychometric function. Consistent with existing theory in the field, the continuous time parametric model of perceptual learning provided a more parsimonious account for the data than the standard nonparametric trial-independent approach. Importantly, our new continuous time parametric model did not do so by producing totally different estimates than the block-based approach. Rather, the fact that the core estimates

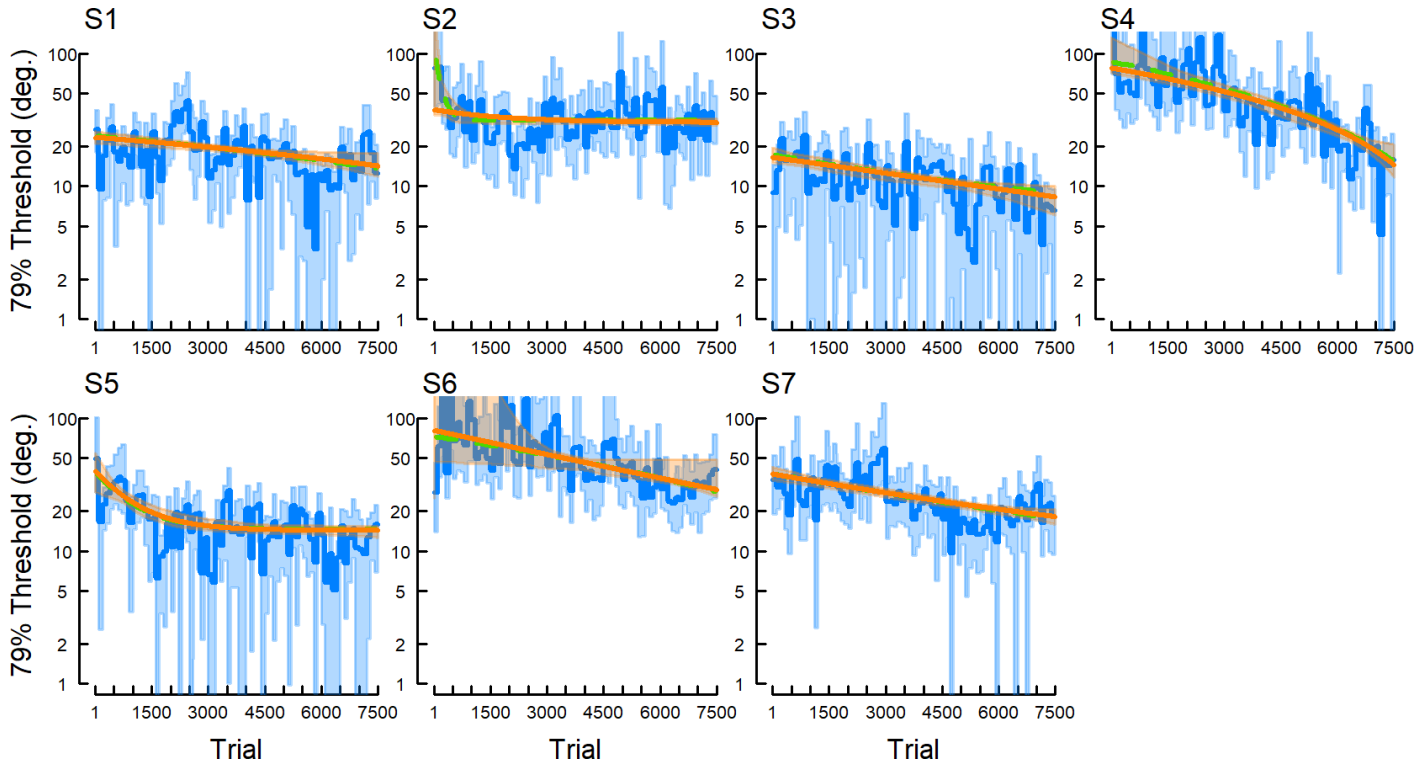


Figure 4. Individual stereo discrimination thresholds based on the 75 separate logistic fits (Equation 1) to 100-trial blocks (block model; blue lines), and the five-parameter continuous model (orange lines). The shaded areas (blue and orange for the respective models) represent bootstrapped 95% confidence intervals for the respective fits. The dashed green line refers to the hybrid model thresholds (see Supplemental Materials).

of thresholds/learning produced by the standard methods were quite similar to those produced by our new approach speaks to the validity of the new approach.

Overfitting

Perhaps the largest difference between the continuous approach and the nonparametric block approach

	Continuous model				Block model			
Subject	*logL*	*AIC*	*BIC*	χ^2	*logL*	*AIC*	*BIC*	χ^2
O1	−788*	1587*	1616*	2015*	−871	1802	1974	5350
O2	−991*	1993*	2021*	1866*	−1104	2267	2439	2570
O3	−918*	1847*	1875*	2127*	−1055	2170	2342	5876
O4	−599*	1207*	1236*	1511*	−2286	4632	4804	25296
O5	−1417*	2845*	2874*	2149*	−1467	2994	3167	2371
O6	−874	1759	1788	6618	−770*	1601*	1773*	4142*
O7	−970*	1950*	1979*	2113*	−1017	2094	2266	2668
S1	−1871*	3752*	3784*	2902*	−3118	6357	6741	4556
S2	−2477*	4965*	4997*	6152*	−2506	5132	5516	6745
S3	−1389*	2788*	2820*	8401*	−1430	2981	3365	9941
S4	−2542	5094	5126*	3713*	−2474*	5068*	5452	3771
S5	−1566*	3142*	3174*	5112*	−1656	3431	3816	8440
S6	−2683	5377	5409*	4020*	−2573*	5265*	5650	4176
S7	−2081*	4172*	4204*	3258*	−3111	6342	6727	4542

Table 3. Extrapolation analysis: The continuous ($np = 5$) and the block model ($np = 20$ and $np = 50$) were fit to early training trials ($nd = 1{,}500$ and $nd = 3{,}000$ of orientation and stereo tasks, respectively), and thresholds were extrapolated by fitting the models to the remaining $nd = 2{,}300$ and $nd = 4{,}500$ orientation and stereo discrimination trials, respectively. Asterisks indicate the model with the better relative fit to the data (based on AIC, BIC, and χ^2).

was in the ability to fit the data without overfitting the data. Indeed, the large number of free parameters resulting from fitting multiple psychometric functions to a large number of blocks makes the nonparametric approach extremely flexible. This flexibility in turn means that the approach will tend to capture any and all fluctuations in the data (we note that this remains true of approaches that fit parametric functions to block threshold estimates; each threshold estimate being fit by the parametric function itself comes from a model fit—it is not in fact "raw data"—and thus such a model still has an extremely large number of free parameters and is exceptionally flexible, though less so than the nonparametric version; see Supplemental Materials for Hybrid model results). This includes fluctuations that arise due to noise alone, which can be substantial in 2-AFC experiments given the standard deviation associated with the binomial distribution. It also includes any number of non-monotonic fluctuations in performance that are unrelated to actual changes in perceptual ability (e.g., mind-wandering/failures of sustained attention, etc.). Data points consistent with these issues can be clearly seen just by simple examination of the block-by-block fits of both experiments, where threshold estimates in contiguous blocks commonly differed by substantial margins (i.e., in a manner that could not possibly be due to a true change in perceptual ability). The parametric continuous model, meanwhile, treats these occasional deviations as noise, and thus their presence did not substantially affect the estimate of participants' true behavioral abilities at the given time points. Although overfitting is a concern any time data is modeled, it is of particular relevance to the domain of perceptual learning. Consider, for example, the simple question of "How much did participants improve?" at a given task. Because the nonparametric block approach tends to overfit noise, this will reduce confidence in early and final performance estimates and thus will reduce confidence in the estimated change between early and final performance (i.e., a few idiosyncratic points in either the first or last block may produce extremely different estimates of total learning).

Trial-specific estimation

In addition to the fact that the continuous time model simply provides a more trustworthy estimate of participant performance and learning, there are a number of other aspects of this method of fitting that may be useful to the field going forward. For instance, one benefit of the continuous parametric model is that it is capable of providing estimates of performance on particular trials of interest. Specifically, in learning experiments, the most critical trials very commonly correspond with the first and last trials of training. Assuming that learning has roughly reached an asymptote by the end of training, the "last trial" estimates given by the continuous parametric approach and nonparametric block approach will tend to be quite similar (excluding issues related to overfitting noise in the block approach). This is because, once participants have hit a rough asymptote, their behavior closely approximates the key block-based assumption that performance is not changing substantially within a block. In contrast though, substantial differences between the nonparametric and parametric approaches are possible in their estimates of early performance. Indeed, given an exponential or power functional form, the early portion of training is when performance changes most rapidly from trial to trial (Badiru, 1992; Dosher & Lu, 2007; Heathcote, Brown, & Mewhort, 2000). Because the nonparametric model aggregates across the first 100 or more trials (Crist et al., 1997; Fahle & Morgan, 1996; Gantz et al., 2007; Z. Liu & Weinshall, 2000; Seitz et al., 2006); with 150, 80, 200, 2,000, and 640 trials, respectively), it is necessarily the case that this approach will collapse over a substantial amount of learning (i.e., the average performance across the block will be substantially better than the performance on the first few trials). Therefore, without using trial-dependent changes in the parameters of the psychometric function, the true "total" amount of learning that has occurred from the first trial of training to the last trial of training will necessarily be underestimated.

Testing functional form of learning

This issue also speaks to another benefit of the continuous parametric model— the ability to more fully examine the functional form of learning. Although our approach here was purely descriptive (i.e., the eventual functional form we chose was simply the one that provided the best overall fits to the data—see Supplemental Materials for other parameterizations; also see Kattner et al., 2017; Snell et al., 2015), the general framework can be easily used to test explicit predictions about the best fitting functional form. This is critical as the observed functional form of learning in a task limits the possible mechanistic models that need to be considered. As such, the question of functional form has thus been investigated in many different fields (Badiru, 1992; Heathcote et al., 2000; Newell & Rosenbloom, 1981) including the field of perceptual learning. For instance, it was reported that the improvements of Vernier acuity thresholds found in adults with amblyopia (not considering slope) across successive blocks (which

were distributed across several days) can be fit by an exponential function (Levi et al., 1997). Similarly, negative exponential functions were used to fit improvements in reading speed in visually impaired patients (Chung, 2011). Other researchers have meanwhile separately modeled improvements in thresholds and slopes (or the width of the psychometric function; Coates & Chung, 2014; Fründ et al., 2011). In seminal work by Dosher and Lu (2007), several different functional learning forms were contrasted, in particular, power versus exponential. The authors found that an exponential functional learning form provided the best fit to individual data (with power only fitting better for the aggregate across participants). Critically though, the authors in this case employed a staircase technique to train participants (140 trials per block). As noted already, because participants are likely to be learning rapidly during this first block, their performance at the end of the block (i.e., which is disproportionately what a staircase analysis focuses on) is almost certainly better than their performance at the beginning of the block. Thus, one likely outcome of this type of trial aggregation approach is to flatten the shape of the learning curve (i.e., by overestimating initial performance), which could in turn potentially affect the best fitting function. Although our experiments were not designed to speak to the exact issue of functional form (e.g., because the participants in our experiments underwent pretesting prior to training, which could also alter estimates of the functional form of learning), the overarching approach could easily be used to examine this question more closely.

The approach can also be extended to address related questions, such as whether the first 100 or 200 trials *should* be included in the full learning curve analysis or whether these trials should be thrown out/treated as practice trials as is common in the literature (i.e., whether there is an early stage of learning that is qualitatively and quantitatively different from the remainder of the learning process). Similarly, the parameterization can be extended or altered to determine whether there are changes in other aspects of the underlying performance functions such as the temporal characteristics of the response biases [e.g., (a) whether it necessary to allow the response bias to change over time or can it be set as a constant—see Supplemental Materials; (b) whether a separate function for the response bias be set for each "day" of the experiment] or the lapse/guess rate (e.g., in our case we assumed a constant lapse rate, but this could also change with training; for examples, see Fründ et al., 2011; Jones, Moore, & Amitay, 2015; Petrov, Dosher, & Lu, 2006) or in the best form of the probability distribution.

Estimating learning and transfer

Further, because the continuous time parametric-model fit has the potential to provide estimates of first trial and last trial performance, it could be additionally useful in designs seeking to address questions regarding total learning (i.e., by comparing first trial and last trial estimates rather than first block and last block), rate of learning, as well as questions related to transfer of learning (Kattner et al., 2017). In the latter cases, "transfer" could be calculated as the difference in performance on the final trial of the training task and the first trial of the transfer task, or the increase in learning rate in the transfer task when compared with the training task. It may also be possible in this endeavor to take advantage of the fact that the parametric model provides for a natural method to extrapolate beyond the data set (i.e., to estimate how performance would have continued to evolve if the participant had carried on with the training task, as compared to how they did perform when asked to perform a new transfer task). Such an extrapolation has no analogue in a nonparametric model that relies on aggregating across trials into blocks, in which performance on the last training block is frequently contrasted with performance on transfer block performance (Liu & Weinshall, 2000).

Finally, we note that the data of primary interest in many perceptual learning papers is not actually the training data, but is instead performance on pre- and post-tests (e.g., to determine whether there are improvements on some untrained task from pretest to posttest). The approach we have outlined here plays an important role in this type of design as well. Learning generalization (i.e., an improvement on an untrained task after training), can take multiple functional forms. There can be an immediate enhancement on the untrained task at posttest (i.e., performance on the first trial of the posttest exceeds performance on the last trial of the pretest). There can also be a change in the rate at which performance improves on the posttest (i.e., performance on the first trial of the posttest matches performance on the last trial of the pretest, but then performance on the post-test rapidly improves). We have referred to these different functional forms of generalization as "transfer" and "learning to learn" respectively (Kattner et al., 2017). Critically, it is easy to misidentify the functional form if a block approach is used to analyze pretest and posttest data (since such an approach does not given an estimate of immediate performance on the posttest nor an estimate of how performance changes throughout the posttest, but instead considers performance during the posttest to be stationary).

Limitations

Although, as we have shown here, modeling learning as a continuous function of experience on a task provides definite benefits in comparison to standard block-by-block analysis of perceptual learning, there are clear limitations to the approach as well. Many of these limitations are related to the strict functional form imposed by the continuous learning function, which would be an issue in circumstances where the data genuinely took a functional form that could not be captured by the model. One circumstance where this would be the case is if there are true discontinuities in learning (Petrov et al., 2005). For instance, in learning experiments that take place over many days, participants may not actually begin each day with the exact same level of performance that they finished with on the previous day (i.e., they may need to readjust to performing the task on each day). This would result in a learning function that is effectively "scalloped," which is a functional form that the current model cannot capture (it would instead tend to smooth over these discontinuities—although the model could certainly be extended in many different ways to account for such data; see, for example, Levi et al., 1997; Yu et al., 2004). Another circumstance that would violate the assumptions of the current approach is if performance improves for a period of time and then proceeds to become worse. This could be the case, for instance, in a long, single-day training experiment where thresholds increase toward the end of training due to fatigue or if participants experience periods of "mind wandering" (McVay & Kane, 2009, 2012); note though that although the nonparametric approach could capture such an occurrence on a block level, it would nonetheless remain difficult to determine how to quantify learning. Finally, the parametric model of perceptual learning is inappropriate for fitting data that is collected via staircase techniques. This is because data collected via staircases (a) rarely includes sufficient spread in stimulus strength to estimate a full psychometric function and (b) what spread in stimulus strength is present is usually highly biased in time, with trials of higher stimulus strength occurring early in the block and trials of lower stimulus strength occurring mainly later in the block (although we note that there are several interesting approaches to examining trial-by-trial data that is generated by staircase techniques, e.g., Ghose, Yang, & Maunsell, 2002; Yang & Maunsell, 2004). There are absolutely a wide range of situations that call for staircase methodology (e.g., when a threshold needs to be estimated very quickly, or when the range of stimulus intensities to present is unknown or could potentially vary substantially between individuals). However, we would suggest that in most learning experiments (which typically involve thousands of trials and where typical performance is usually reasonably well known) staircases could easily be replaced by methods that are more amenable to the analysis techniques described here. These methods could be augmented by using the technique to adapt level of difficulty so as to maximize learning.

Indeed, it is certainly the case that, given the analysis methods developed and demonstrated here, the methodology that was used to produce the current data can be significantly improved going forward. Specifically, stimulus strength in all cases was drawn from a static uniform range throughout training. This then necessarily meant that as learning proceeded, more and more of the range resulted in ceiling level performance, which is neither ideal with respect to measuring ability (as the far ceiling part of the curve does not help constrain the psychometric function) nor with respect to producing learning (although there is certainly virtue to having some easy trials, it is generally accepted that learning is most efficient when the task is challenging, but doable—when errors are being made, but these errors are informative; Chu, Dosher, & Lu, 2010). It is thus possible that there could be virtue to using the current completely offline analysis approach in an online manner to control the range over which stimulus strength is sampled. This would, in some ways, be the best of both worlds—the uniform random sampling of stimulus strengths allows for continuous estimates of learning—and drifting that range provides a natural way to keep difficulty level constant throughout training (without having to rely on staircases).

Conclusions

Here, we have presented a method to parametrically fit the changes in the psychometric function occurring in perceptual learning studies. Future work with larger datasets will be necessary to identify the best parameterizations of both learning functions and the psychometric functions (and the correct parameterizations may differ across learning domains). Indeed, we note that we ourselves have used a slightly different parameterization in our previous empirical work (Green et al., 2015; Kattner et al., 2017), which used even fewer free parameters, at the cost of additional flexibility. The best tradeoff between these is thus also to be determined. Furthermore, it is likely the case that this approach will be fruitful for analyzing perceptual learning experiments that use alternative designs—such as target present/target absent designs (Ahissar & Hochstein, 1997); however, the parameterization will similarly need to be altered to account for the difference in design. In all, the current data clearly demonstrates that continuous parametric fitting is a flexible tool that

allows for these comparisons to be made using few free parameters and without assuming within-block stationarity.

Keywords: psychometric function, perceptual learning, parametric model

Acknowledgment

This work was supported by the Office of Naval Research grant ONR - N000141712049.

Commercial relationships: none.
Corresponding author: C. Shawn Green.
Email: cshawn.green@wisc.edu.
Address: Department of Psychology, University of Wisconsin-Madison, Madison, WI, USA.

References

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*(6631), 401–406. http://doi.org/10.1038/387401a0

Astle, A. T., Blighe, A. J., Webb, B. S., & McGraw, P. V. (2015). The effect of normal aging and age-related macular degeneration on perceptual learning. *Journal of Vision*, *15*(10):16, 1–16, doi:10.1167/15.10.16. [PubMed] [Article]

Badiru, A. B. (1992). Computational survey of univariate and multivariate learning curve models. *IEEE Transactions on Engineering Management*, *39*(2), 176–188. http://doi.org/10.1109/17.141275

Ball, K., & Sekuler, R. (1987). Direction-specific improvement in motion discrimination. *Vision Research*, *27*(6), 953–965. http://doi.org/10.1016/0042-6989(87)90011-3

Beard, B. L., Levi, D. M., & Reich, L. N. (1995). Perceptual learning in parafoveal vision. *Vision Research*, *35*(12), 1679–1690. http://doi.org/10.1016/0042-6989(94)00267-P

Bejjanki, V. R., Beck, J. M., Lu, Z.-L. L., & Pouget, A. (2011). Perceptual learning as improved probabilistic inference in early sensory areas. *Nature Neuroscience*, *14*(5), 642–648. http://doi.org/10.1038/nn.2796

Casey, M. C., & Sowden, P. T. (2012). Modeling learned categorical perception in human vision. *Neural Networks*, *33*, 114–126. http://doi.org/10.1016/j.neunet.2012.05.001

Chu, W., Dosher, B., & Lu, Z.-L. (2010). The rate of perceptual learning at a fixed accuracy threshold is improved by feedback and by mixture with easier trials. *Journal of Vision*, *9*(8): 882, doi:10.1167/9.8.882. [Abstract]

Chung, S. T. L. (2011). Improving reading speed for people with central vision loss through perceptual learning. *Investigative Opthalmology & Visual Science*, *52*(2), 1164–1170. [PubMed] [Article]

Coates, D. R., & Chung, S. T. L. (2014). Changes across the psychometric function following perceptual learning of an RSVP reading task. *Frontiers in Psychology*, *5*(DEC). http://doi.org/10.3389/fpsyg.2014.01434

Crist, R. E., Kapadia, M., Westheimer, G., & Gilbert, C. D. (1997). Perceptual learning of spatial localization: Specificity for orientation, position and contextocalization: Specificity for orientation, position and context. *Journal of Neurophysiology*, *78*(6), 2889–2894.

Dosher, B. A., & Lu, Z.-L. (1998). Perceptual learning reflects external noise filtering and internal noise reduction through channel reweighting. *Proceedings of the National Academy of Sciences, USA*, *95*(23), 13988–13993. http://doi.org/10.1073/pnas.95.23.13988

Dosher, B. A., & Lu, Z.-L. (2007). The functional form of performance improvements in perceptual learning: Learning rates and transfer. *Psychological Science*, *18*(6), 531–539. http://doi.org/10.1111/j.1467-9280.2007.01934.x

Dosher, B. A., & Lu, Z.-L. L. (2000). Mechanisms of perceptual attention in precuing of location. *Vision Research*, *40*(10–12), 1269–1292. http://doi.org/10.1016/S0042-6989(00)00019-5

Fahle, M., & Edelman, S. (1993). Long-term learning in vernier acuity: Effects of stimulus orientation, range and of feedback. *Vision Research*, *33*(3), 397–412. http://doi.org/10.1016/0042-6989(93)90094-D

Fahle, M., & Morgan, M. (1996). No transfer of perceptual learning between similar stimuli in the same retinal position. *Current Biology*, *6*(3), 292–297. http://doi.org/10.1016/S0960-9822(02)00479-7

Fendick, M., & Westheimer, G. (1983). Effects of practice and the separation of test targets on foveal and peripheral stereoacuity. *Vision Research*, *23*(2), 145–150. http://doi.org/10.1016/0042-6989(83)90137-2

Fründ, I., Haenel, N. V., & Wichmann, F. A. (2011). Inference for psychometric functions in the presence of nonstationary behavior. *Journal of Vision*, *11*(6):16, 1–19, doi:10.1167/11.6.16. [PubMed] [Article]

Gantz, L., Patel, S. S., Chung, S. T. L., & Harwerth, R.

S. (2007). Mechanisms of perceptual learning of depth discrimination in random dot stereograms. *Vision Research*, *47*(16), 2170–2178. http://doi.org/10.1016/j.visres.2007.04.014

Ghose, G. M., Yang, T., & Maunsell, J. H. R. (2002). Physiological correlates of perceptual learning in monkey V1 and V2. *Journal of Neurophysiology*, *87*(4), 1867–1888. http://doi.org/10.1152/jn.00690.2001

Green, C. S., Kattner, F., Siegel, M. H., Kersten, D., & Schrater, P. R. (2015). Differences in perceptual learning transfer as a function of training task. *Journal of Vision*, *15*(10):5, 1–14, doi:10.1167/15.10.5. [PubMed] [Article]

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*(3), 280–301. http://doi.org/10.1016/j.bandl.2005.06.001

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207. http://doi.org/10.3758/BF03212979

Herzog, M. H., & Fahle, M. (1997). The role of feedback in learning a vernier discrimination task. *Vision Research*, *37*(15), 2133–2141.

Herzog, M. H., & Fahle, M. (1998). Modeling perceptual learning: Difficulties and how they can be overcome. *Biological Cybernetics*, *78*(2), 107–117. http://doi.org/10.1007/s004220050418

Herzog, M. H., & Fahle, M. (1999). Effects of biased feedback on learning and deciding in a vernier discrimination task, *Vision Research*, *39*, 4232–4243.

Jacobs, R. A., & Kruschke, J. K. (2011). Bayesian learning theory applied to human cognition. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(1), 8–21. http://doi.org/10.1002/wcs.80

Jones, P. R., Moore, D. R., & Amitay, S. (2015). The role of response bias in perceptual learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *2015*(5), 1456–1470. http://doi.org/10.1037/xlm0000111

Kattner, F., Cochrane, A., Cox, C. R., Gorman, T. E., & Green, C. S. (2017). Perceptual learning generalization from sequential perceptual training as a change in learning rate. *Current Biology*, *27*(6), 840–846. http://doi.org/10.1016/j.cub.2017.01.046

Klein, S. A. (2001). Measuring, estimating, and understanding the psychometric function: A commentary. *Perception & Psychophysics*, *63*(8), 1421–1455. http://doi.org/10.3758/BF03194552

Law, C.-T., & Gold, J. I. (2009). Reinforcement learning can account for associative and perceptual learning on a visual-decision task. *Nature Neuroscience*, *12*(5), 655–663. http://doi.org/10.1038/nn.2304

Levi, D. M., Polat, U., & Hu, Y. S. (1997). Improvement in Vernier acuity in adults with amblyopia: Practice makes better. *Investigative Ophthalmology and Visual Science*, *38*(8), 1493–1510. [PubMed] [Article]

Liu, J., Dosher, B. A., & Lu, Z.-L. (2014). Modeling trial by trial and block feedback in perceptual learning. *Vision Research*, *99*, 46–56. http://doi.org/10.1016/j.visres.2014.01.001

Liu, Z., & Weinshall, D. (2000). Mechanisms of generalization in perceptual learning. *Vision Research*, *40*(1), 97–109. http://doi.org/10.1016/S0042-6989(99)00140-6

Lu, Z.-L., Hua, T., Huang, C.-B., Zhou, Y., & Dosher, B. A. (2011). Visual perceptual learning. *Neurobiology of Learning and Memory*, *95*(2), 145–151. http://doi.org/10.1016/j.nlm.2010.09.010

Matthews, N., Liu, Z., Geesaman, B. J., & Qian, N. (1999). Perceptual learning on orientation and direction discrimination. *Vision Research*, *39*(22), 3692–3701.

Mazur, J. E., & Hastie, R. (1978). Learning as accumulation: A reexamination of the learning curve. *Psychological Bulletin*, *85*(6), 1256–1274. http://doi.org/10.1037/0033-2909.85.6.1256

McVay, J. C., & Kane, M. J. (2009). Conducting the train of thought: Working memory capacity, goal neglect, and mind wandering in an executive-control task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *35*(1), 196–204. http://doi.org/10.1037/a0014104

McVay, J. C., & Kane, M. J. (2012). Drifting from slow to "d'oh!": Working memory capacity and mind wandering predict extreme reaction times and executive control errors. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(3), 525–549. http://doi.org/10.1037/a0025896

Michel, M. M., & Jacobs, R. A. (2007). Parameter learning but not structure learning: A Bayesian network model of constraints on early perceptual learning. *Journal of Vision*, *7*(1):4, 1–18, doi:10.1167/7.1.4. [PubMed] [Article]

Moscatelli, A., Mezzetti, M., & Lacquaniti, F. (2012). Modeling psychophysical data at the population-level: The generalized linear mixed model. *Journal of Vision*, *12*(11):26, 1–17, doi:10.1167/12.11.26. [PubMed] [Article]

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms

of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–51). Hillsdale, NJ: Lawrence Erlbaum.

Petrov, A. A., Dosher, B. A., & Lu, Z.-L. (2005). The dynamics of perceptual learning: An incremental reweighting model. *Psychological Review*, *112*(4), 715–743. http://doi.org/10.1037/0033-295X.112.4.715

Petrov, A. A., Dosher, B. A., & Lu, Z. L. (2006). Perceptual learning without feedback in non-stationary contexts: Data and model. *Vision Research*, *46*(19), 3177–3197. http://doi.org/10.1016/j.visres.2006.03.022

Poggio, T., Fahle, M., & Edelman, S. (1992). Fast perceptual learning in visual hyperacuity. *Science*, *256*(5059), 1018–1021. http://doi.org/10.1126/science.1589770

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. http://doi.org/10.1037/h0042519

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(6088), 533–536. http://doi.org/10.1038/323533a0

Seitz, A. R., Nanez, J. E., Holloway, S., Tsushima, Y., & Watanabe, T. (2006). Two cases requiring external reinforcement in perceptual learning. *Journal of Vision*, *6*(9):9, 966–973. [PubMed] [Article]

Snell, N., Kattner, F., Rokers, B., & Green, C. S. (2015). Orientation transfer in vernier and stereoacuity training. *PLoS One*, *10*(12),e0145770. http://doi.org/10.1371/journal.pone.0145770

Sotiropoulos, G., Seitz, A. R., & Seris, P. (2011). Changing expectations about speed alters perceived motion direction. *Current Biology*, *21*(21), R883–R884. http://doi.org/10.1016/j.cub.2011.09.013

Spratling, M. W., & Johnson, M. H. (2001). Dendritic inhibition enhances neural coding properties. *Cerebral Cortex*, *11*(12), 1144–1149. http://doi.org/10.1093/cercor/11.12.1144

Spratling, M. W., & Johnson, M. H. (2006). A feedback model of perceptual learning and categorization. *Visual Cognition*, *13*(2), 129–165. http://doi.org/10.1080/13506280500168562

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, *35*(17), 2503–2522. http://doi.org/10.1016/0042-6989(95)00016-X

Vaina, L. M., Sundareswaran, V., & Harris, J. G. (1995). Learning to ignore: Psychophysics and computational modeling of fast learning of direction in noisy motion stimuli. *Cognitive Brain Research*, *2*(3), 155–163. http://doi.org/10.1016/0926-6410(95)90004-7

Yang, T., & Maunsell, J. H. R. (2004). The effect of perceptual learning on neuronal responses in monkey visual area V4. *The Journal of Neuroscience*, *24*(7), 1617–1626. http://doi.org/10.1523/JNEUROSCI.4442-03.2004

Yu, C., Klein, S. A., & Levi, D. M. (2004). Perceptual learning in contrast discrimination and the (minimal) role of context. *Journal of Vision*, *4*(3):4, 169–182, doi:10.1167/4.3.4. [PubMed] [Article]

Zhaoping, L., Herzog, M. H., & Dayan, P. (2003). Nonlinear ideal observation and recurrent preprocessing in perceptual learning. *Network (Bristol, England)*, *14*(2), 233–247. http://doi.org/10.1088/0954-898X/14/2/304

Appendix 3. Kattner, Cochrane, Cox, Gorman, and Green (2017)

Report

Current Biology

Perceptual Learning Generalization from Sequential Perceptual Training as a Change in Learning Rate

Highlights

- Training on multiple perceptual tasks produced significant learning generalization
- Learning generalization manifested as increased learning rate ("learning to learn")
- Standard methodology would tend to miss or misidentify this type of generalization

Authors

Florian Kattner, Aaron Cochrane, Christopher R. Cox, Thomas E. Gorman, C. Shawn Green

Correspondence

cshawn.green@wisc.edu

In Brief

The extent to which learning generalizes to new tasks is a key concern in the study of perceptual learning. Kattner, Cochrane, et al. report three experiments demonstrating that training on a series of tasks can induce generalization that manifests only in terms of increases in learning rate ("learning to learn"), not immediate performance.

Kattner et al., 2017, Current Biology *27*, 840–846
March 20, 2017
http://dx.doi.org/10.1016/j.cub.2017.01.046

Perceptual Learning Generalization from Sequential Perceptual Training as a Change in Learning Rate

Florian Kattner,[1,3] Aaron Cochrane,[2,3] Christopher R. Cox,[2] Thomas E. Gorman,[2] and C. Shawn Green[2,4,*]
[1]Institute of Psychology, Technische Universität Darmstadt, Alexanderstr. 10, 64283 Darmstadt, Germany
[2]Department of Psychology, University of Wisconsin–Madison, 1202 West Johnson Street, Madison, WI 53706-1611, USA
[3]Co-first author
[4]Lead Contact
*Correspondence: cshawn.green@wisc.edu
http://dx.doi.org/10.1016/j.cub.2017.01.046

SUMMARY

With practice, humans tend to improve their performance on most tasks. But do such improvements then generalize to new tasks? Although early work documented primarily task-specific learning outcomes in the domain of perceptual learning [1–3], an emerging body of research has shown that significant learning generalization is possible under some training conditions [4–9]. Interestingly, however, research in this vein has focused nearly exclusively on just one possible manifestation of learning generalization, wherein training on one task produces an immediate boost to performance on the new task. For instance, it is this form of generalization that is most frequently referred to when discussing learning "transfer" [10, 11]. Essentially no work in this domain has focused on a second possible manifestation of generalization, wherein the knowledge or skills acquired via training, despite not being directly applicable to the new task, nonetheless allow the new task to be learned more efficiently [12–15]. Here, in both the visual category learning and visual perceptual learning domains, we demonstrate that sequentially training participants on tasks that share a common high-level task structure can produce faster learning of new tasks, even in cases where there is no immediate benefit to performance on the new tasks. We further show that methods commonly employed in the field may fail to detect or else conflate generalization that manifests as increased learning rate with generalization that manifests as immediate boosts to performance. These results thus lay the foundation for the various routes to learning generalization to be more thoroughly explored

RESULTS

Experiment 1: Generalization as a Change in Learning Rate in Novel-Shape Categorization

We have previously identified multiple distinct mechanisms that could, in principle, promote increases in the rate at which new tasks are learned (which has been referred to as the "learning to learn" form of generalization) without engendering any immediate benefit to performance [15–17]. In the present paper, we chose to focus on one of these mechanisms, which we have referred to as the "knowledge-based" mechanism. Here, through exposure to many different tasks that all share some higher-level structure or components, participants could potentially learn those regularities that exist across the individual tasks [14, 18–20]. Importantly, knowledge of such higher-level regularities need not provide any direct insight regarding how one should interpret or respond to stimuli when beginning a new task. Instead, this knowledge may serve to constrain or order the to-be-learned task space. If this is the case, such training will produce faster learning of new tasks without any immediate benefit to performance.

As an initial demonstration of the conditions that promote this form of generalization, we chose a domain where it is reasonably straightforward to induce the necessary higher-level shared task structure. While there are a number of domains where this would be possible, we selected novel-shape categorization. This choice was based primarily on the robust body of work outlining clear similarities between the novel-shape categorization domain and the perceptual learning domain, suggesting that the former could be a good model for the latter [21–26].

We first created a continuous 2D space from which individual novel shapes could be drawn. We then defined eight unique categories within that space, pairs of which could be utilized in four separate categorization learning tasks (Figure 1A; Supplemental Experimental Procedures). Importantly, while the individual shapes, categories, and category boundaries were unique to each learning task, many other higher-order features were shared across learning tasks. For instance, although the categories were placed in different parts of the 2D space in the different learning tasks, the general shape of the categories was shared (i.e., 2D Gaussians with similar parameters). Other shared aspects included the fact that the two to-be-discriminated categories in each learning task were always linearly separable and, furthermore, were always separable along a single dimension. Critically, although this shared structure provided no information that would be immediately applicable in a new task, it should nonetheless allow new tasks that share this same structure to be learned more quickly.

Twenty-four participants underwent this series of four novel-shape categorization learning tasks (60 trials each). In examining their behavior, we first assessed whether any immediate

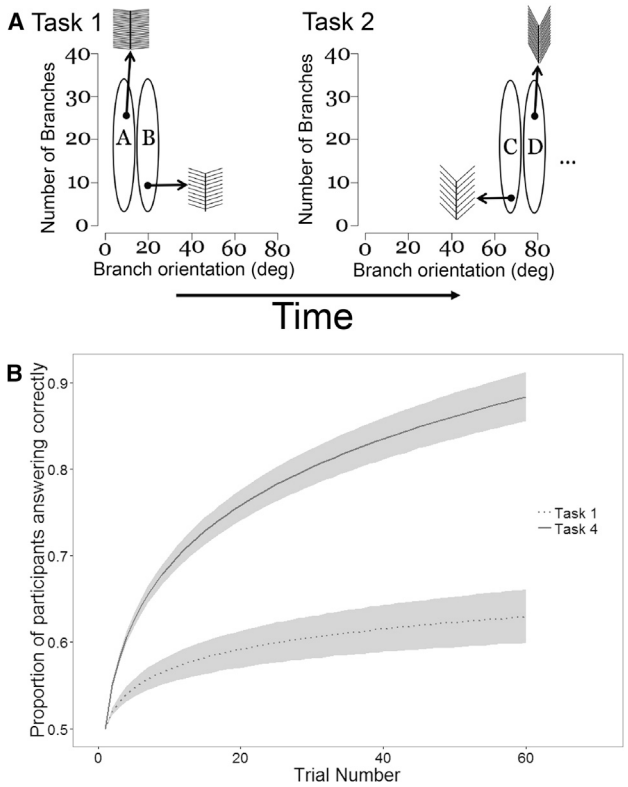


Figure 1. Sequential Novel Shape Categorization Task and Results

(A) Novel shapes ("feathers") were drawn from a 2D space, where one dimension corresponded to the number of branches on the feather and the other dimension corresponded to the orientation of the branches (e.g., a point in the top left of this space produces a feather with many branches with a very flat orientation, while a point in the bottom right of this space produces a feather with few branches with a very steep orientation). For each learning task, feathers were drawn from one of two categories defined by 2D Gaussians in the space (e.g., category A versus category B in the left panel and category C versus category D in the right panel). While the various tasks involve totally different shapes, categories, and category boundaries, they share a certain degree of high-level structure, including the general shape of the categories in the space and the fact that the discriminant that best separates the categories lies along a single dimension.

(B) Although participants started with the same (chance-level) performance in both their first and final categorization tasks, they learned much more quickly in the final categorization task as compared to the first (note that, for correspondence with experiment 2, we plotted only the first and final categorization tasks here; see Figure S1 for behavioral performance across all four categorization tasks and Figure S2 for the bootstrapped learning slope estimates across all four categorization tasks). Error bands represent 95% confidence intervals.

changes in performance were observed from task to task by examining first-trial performance across the four tasks. Consistent with our expectation that no such immediate changes would be observed, participants began with similar, chance-level performance in all four tasks (task 1: M = 0.42; task 2: M = 0.54; task 3: M = 0.54; task 4: M = 0.46; none of these values were significantly different either from one another or from chance; Figures 1B and S1). We next tested the hypothesis that participants would learn more quickly as they progressed from task to task. To this end, we computed bootstrapped estimates of the learning curve for each task (see the Supplemental Experimental Procedures) and then assessed the extent to which the rate of learning was increasing from task to task. Consistent with our hypothesis, we found that participants did indeed learn to categorize more quickly as they moved from task to task (e.g., the slope of the average best fitting line to the learning rates across tasks was significantly above zero, indicating faster learning from task to task: $b = 0.026$, $p < 0.001$; see Figures 1B, S1, and S2 and the Supplemental Experimental Procedures for additional quantification). These results thus strongly indicate that properly sequenced training on tasks containing shared higher-order structure can induce generalization in the form of learning rate, in the absence of any immediate changes in performance.

Experiment 2: Generalization as a Change in Learning Rate in Perceptual Learning

Given the results described above, we next applied similar logic to the domain of perceptual learning. Specifically, we first designed a set of five perceptual learning tasks—visual grating spatial frequency categorization, color lightness categorization, dot bisection, Gabor orientation categorization, and dot motion direction categorization. These tasks shared little in terms of the base features of the stimuli utilized in the various tasks (e.g., the stimuli included visual gratings of various spatial frequencies, square color patches of various lightness, three roughly vertically arranged dots, visual Gabors of various orientations, and fields of dots moving in various average directions; some stimuli were presented centrally, and some stimuli were presented peripherally at different spatial positions depending on the task; see the Supplemental Experimental Procedures). Note that the particular base features were chosen partially because they are dimensions along which learning specificity has been commonly observed in the perceptual learning literature (i.e., there are many examples of learning that failed to generalize across position, spatial frequency, orientation, motion direction, etc.; for a review, see [27]).

Despite the lack of similarity at the level of exact stimuli, the tasks in experiment 2 nonetheless shared a great deal of higher-level structure. For instance, the base timing structure, including a 150-ms stimulus presentation followed by a 500-ms mask, was shared across all tasks. Other higher-level aspects that were shared across tasks included the fact that stimuli were always drawn from a uniform distribution and that the category boundary was always found in the center of that uniform distribution. As was true in experiment 1, the shared structure across tasks thus provided no information regarding the exact choice that should be made on the first trial of a new task (e.g., none of the shared structure indicated whether an observed Gabor was clockwise or counterclockwise relative to a given reference angle). However, the shared structure could, for instance, provide information that would allow the participants to more quickly learn to separate signal from noise (and thus improve performance more quickly overall).

Thirteen participants were trained sequentially on the five different perceptual learning tasks (first four tasks: 800 trials per day, 2 days each; fifth task: 500 trials on a single day). Meanwhile, a second cohort of ten participants underwent only the final task (note that we refer to the comparison between these

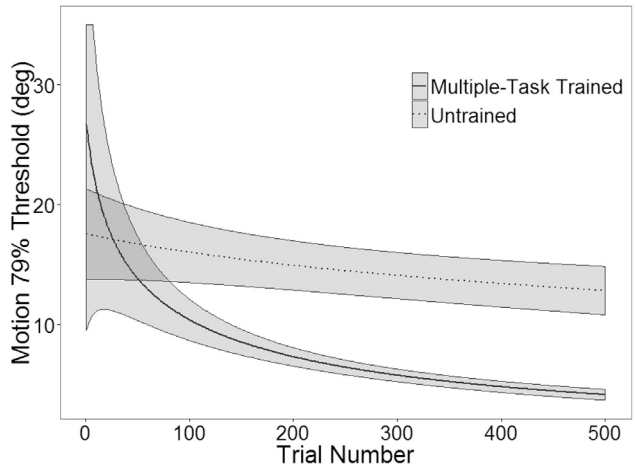


Figure 2. "Learning to Learn" without "Transfer" in Perceptual Learning

Although both the multiple-task trained and untrained participants started with identical initial levels of performance on the dot motion direction categorization task, the trained individuals (i.e., those participants who had previously undergone perceptual learning on four tasks with similar high-level structure) learned much faster (see Figure S4 for fitting method comparison and individual-level data, Figure S3 for the first/fourth comparison, and Table S1 for details on the second and third trained tasks). Error bands represent 95% confidence intervals.

two groups as the "multiple-task trained/untrained comparison"). This setup allowed us to then directly assess whether the multiple-task trained participants showed differences in either initial performance or learning rate on the final perceptual learning task as compared to the untrained participants. Furthermore, in order to allow for additional, within-participant tests of our hypotheses, six of the multiple-task trained participants performed the first four perceptual learning tasks in one order, while the remaining seven performed the same tasks in the reverse order (see the Supplemental Experimental Procedures). By combining performance (via *Z* scoring) across the participants' respective first training tasks (six participants: spatial frequency; seven participants: orientation) and the participants' respective fourth training tasks (vice versa), we could make a similar set of comparisons as in the multiple-task trained/untrained case but within participants rather than between participants (we refer to this as the "first/fourth comparison" below).

For each participant and task, the data were fit via a time-evolving logistic function that has previously been used by our group to examine learning in the perceptual domain [7, 8]. Critically, unlike standard fitting techniques that aggregate over large blocks of trials, our statistical approach allows for an estimate both of immediate changes in performance (making use of the estimated threshold on the first trial of each new task) and of changes in learning rate (making use of the rate at which the psychometric function changes over time). Our prediction was that for both the multiple-task trained/untrained and first/fourth comparisons, we would see no differences in first-trial performance but that we would see significant differences in learning rate.

As can be seen in Figure 2 (see also Figure S3), both hypotheses were confirmed. No significant difference in first-trial performance was seen in either the multiple-task trained/untrained comparison (untrained: threshold = 17.55 ± 1.88; multiple-task trained: threshold = 26.71 ± 8.61; t(13.1) = 1.04, p = 0.32) or the first task/fourth task comparison (first task: *Z*-scored threshold = 1.36 ± 0.38; fourth task: *Z*-scored threshold = 0.85 ± 0.25; $t(12) = 0.99$, p = 0.34). There was, however, a clear difference in the rate at which participants learned in both the multiple-task trained/untrained comparison (rate of change for the psychometric function: for untrained, $5.29 \times 10^{-5} \pm 1.60 \times 10^{-5}$; for multiple-task trained, $4.85 \times 10^{-4} \pm 6.06 \times 10^{-5}$; $t(13.7) = 6.9$; $p < 0.001$; see the Supplemental Experimental Procedures) and the first/fourth comparison (*z*-scored rate of change for the psychometric functions: for first tasks, −0.59 ± 0.13; for fourth tasks, 0.59 ± 0.24; $t(12) = 5.7$, $p < 0.001$). Therefore, as was true in experiment 1, the data clearly indicate that sequential training on multiple perceptual learning tasks that share higher-order structure can induce changes in learning rate in the absence of any immediate changes in performance.

Experiment 3: Assessing the Role of Training Variety

Two key questions are raised by the results of experiment 2. The first question is whether the results were dependent upon a training regimen that included multiple tasks or whether the same effect would be induced by training on a single task for the same total amount of time. Many theoretical frameworks [15, 28], suggest a critical role for variety of experience, as encountering the same higher-level structure in multiple different tasks/contexts is a cue that the structure is indeed more broadly applicable. However, it is certainly possible that experience with the same statistical structure in a single exemplar task would be equivalently valuable.

The second question is whether the observed enhancement in the learning of new tasks is indeed dependent upon shared statistical structure. The results of experiment 2 clearly show that after learning a number of tasks that share the same structure, participants are able to learn a new task that shares that same learned structure more quickly. However, an additional explicit prediction from our framework is that if this learned structure is violated in a new task, then performance should suffer (i.e., would be worse than if participants hadn't completed any previous learning tasks).

To address these issues, we again trained nine new participants sequentially on five different perceptual learning tasks. Unlike in experiment 2, though, these participants began by completing a total of 6,400 trials of the initial orientation training task (800 trials per day, 8 days; i.e., the same number of total trials/days as for the first four training tasks in experiment 2; as such, this group is referred to as a the "single-task trained group"; see the Supplemental Experimental Procedures). The participants then completed 800 trials of the same motion task as in experiment 2. Comparing the first 500 trials of motion task performance of the participants in experiment 3 with those in experiment 2 thus offers a clear assessment of the role of variety in the "learning to learn" effect (i.e., both groups would have completed 6,400 trials of perceptual learning prior to completing the motion learning task— experiment 2 participants having done so across four tasks and experiment 3 participants having done so across a single task). After the motion task, participants completed the same basic color lightness categorization and dot bisection tasks from experiment 2. This was done in order to

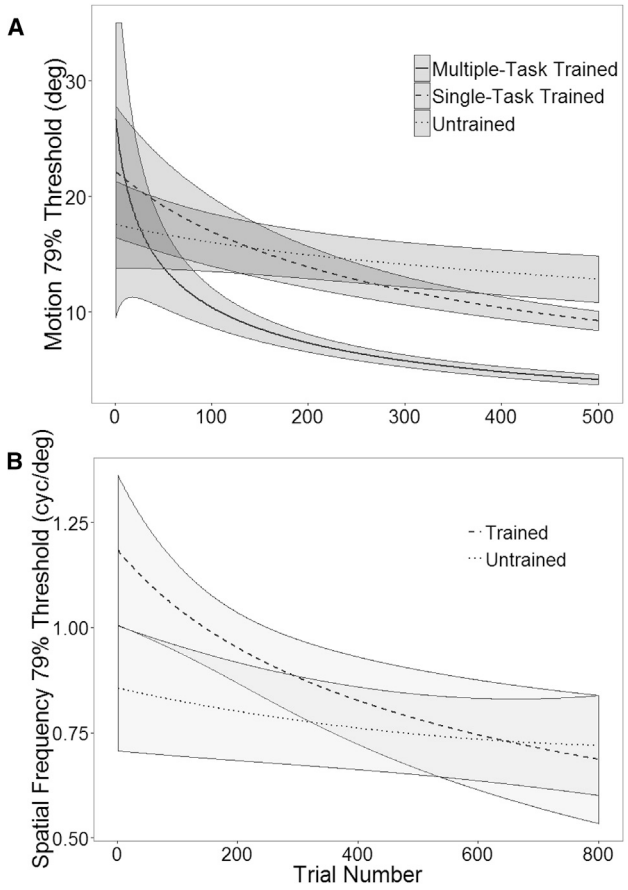


Figure 3. Assessing the Role of Training Variety and Violations of Task Structure

(A) The performance of the single-task trained group in experiment 3 on the motion learning task was intermediate to both of the groups from experiment 2. Learning was significantly faster than that of the untrained group but slower than that of the multiple-task trained group. Error bands represent 95% confidence intervals.

(B) The violations of the learned-task structure in the spatial frequency learning task resulted in poorer initial performance in the trained group of experiment 3 when compared to the untrained group. However, there was a trend for the learning rate to still be faster in the trained group than in the untrained group. This would be consistent with the fact that, while one aspect of the training was violated in the spatial frequency task (i.e., the exact temporal order/structure), many other aspects remained shared (e.g., the fact that some aspects of the presentation were stimuli, whereas some were noise; the fact that the stimulus was quickly presented; the fact that the stimuli differed along a single continuous dimension with the category boundary lying in the center of the uniform distribution; etc.). Error bands represent 95% confidence intervals.

ensure that the trained participants in experiment 3 experienced the higher-level task structure with the same amount of variety as the participants in experiment 2 before completing a final task aimed at addressing the second key question above (i.e., whether the enhanced learning was dependent upon new tasks that share the same structure as learned tasks). To this end, as a final task, participants completed a new version of the spatial frequency task that was redesigned from the experiment 2 version to violate one major aspect of the previous training. Specifically, in this new version, the random 500-ms mask came before the stimulus, rather than after. If one key piece of information taught by the preceding training tasks was the temporal relationship between stimulus (target) and noise (mask), this final task would be expected to be more difficult for the participants who had undergone the training than for another new group of participants (n = 9), who only performed the spatial frequency task.

Answering the question of the role of variety, the learning rate parameters in the motion learning task were significantly different between the single-task trained group from experiment 3 and both the untrained group from experiment 2 and the multiple-task trained group from experiment 2 (Figure 3A). Specifically, the single-task trained group learned significantly more quickly than the untrained group (for the single-task trained group, $1.59 \times 10^{-4} \pm 1.62 \times 10^{-5}$; for the untrained group, $5.29 \times 10^{-5} \pm 1.60 \times 10^{-5}$; $t(16.9) = 4.66$; $p < 0.001$) but significantly more slowly than the multiple-task trained group (for the multiple-task trained group, $4.85 \times 10^{-4} \pm 6.06 \times 10^{-5}$; $t(13.7) = 5.2$, $p < 0.001$).

For the question of the effect of violated task structure, as expected, repeated training on tasks with similar statistics led to a decreased initial performance when these statistics were violated (Figure 3B). Single-task trained participants had significantly higher initial spatial frequency thresholds than participants who had not completed any training (we use the label "single-task trained" in order to differentiate the trained group in experiment 3 from the trained group in experiment 2; however, note again that the single-task trained group had in fact been trained on four separate tasks before they were trained on the spatial frequency task: for single-task trained, 1.18 ± 0 0.089; for untrained, 0.86 ± 0 0.074; $t(16.1) = 2.8$; $p = 0.012$). As shown in Figure 3B, this difference in initial threshold was accompanied by a non-significantly faster learning rate for single-task trained versus untrained participants (for single-task trained, $1.21 \times 10^{-3} \pm 4.03 \times 10^{-4}$; for untrained, $3.74 \times 10^{-4} \pm 2.46 \times 10^{-4}$; $t(13.4) = 1.8$, $p = 0.098$).

While the results of both questions are broadly consistent with the expectations of the our theoretical framework, caution is warranted when attempting to interpret the exact changes that gave rise to the final learning task performance. For instance, one interpretation of the motion learning task data in experiment 3 involves only changes in a positive direction—that is, that the single-task training led to some learning of the shared structure, but it was of a lesser degree than the learning induced by the multiple-task training. An alternative interpretation, though, would involve changes in both a positive direction (some learning of the shared structure) and a negative direction (i.e., stronger learning of the non-shared task structure, such as the spatial locations that attention should be guided toward). Similarly, while initial performance on the spatial frequency learning task in experiment 3 was worse for trained than for untrained participants (consistent with violated expectations), there was nonetheless a strong trend toward faster learning (that would be consistent with the fact that, although one task aspect was violated in the spatial frequency task—i.e., the exact timing—many other task aspects were still shared—e.g., the fact that critical information was briefly presented, that some of the presentation was pure noise, etc.). This is particularly critical, as it is unclear that "unlearning" previously learned information should follow the same temporal dynamics (but in an opposite direction) as the initial learning [29].

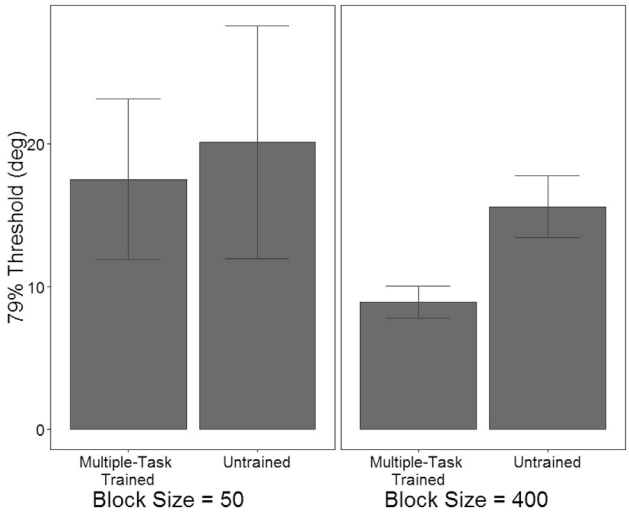


Figure 4. Inferences about Generalization that Would Have Been Drawn via More Standard Techniques

Although the data analysis technique we employed in experiments 2 and 3 specifically models changes in performance as a continuous function of time, it is considerably more standard in the literature to fit data as a single block. However, if this block is too small (left bars), no difference between groups may be detected (i.e., differences that would have been present due to "learning to learn" would be missed). Conversely, if this block is too large (right bars), generalization will be detected, but it will be wrongly identified as "transfer" rather than "learning to learn" (i.e., without modeling performance as a function of time, there is no way to determine whether the observed differences were present immediately or evolved through time). See also Figure S4 for a comparison of fitting methods. Error bars represent 95% confidence intervals.

DISCUSSION

The results of the present investigation clearly demonstrate that properly designed sequential training can induce perceptual learning generalization that manifests as a change in learning rate, in the absence of any immediate changes in performance. This is consistent with the broad idea that learning higher-level structure can, in turn, facilitate learning the individual parameters of new tasks, thus inducing what has been called, in various parts of the literature, "learning to learn" [14–17, 20, 28, 30–33]. Critically, this is not a simple matter of directly applying some known information to a new task (either immediately or delayed through time), which is commonly referred to as "transfer" of learning. Indeed, in our case here, the higher-level structure that exists across tasks (e.g., the consistent timing information) provides no information that would directly inform actions in each new task (i.e., the timing information provides no information regarding what separates "high" from "low" spatial frequency responses).

In order to explore this distinction further, significant methodological changes may be necessary for the perceptual learning field going forward. For instance, one of the more common designs used to examine perceptual learning generalization involves training on some perceptual learning task "A" followed by a single block of some generalization task "B." Unfortunately, this type of design may result in the "learning to learn" form of generalization being missed entirely, or else it may result in the "learning to learn" form of generalization being mislabeled as immediate transfer (depending on the number of trials used and the rate at which the task is learned). To make this issue explicit, we assessed the inferences that would have arisen had we utilized more typical methodological and statistical approaches in experiment 2. In the first case, we fit the data for the first 50 trials of the generalization tasks (mimicking a very short generalization task), while in the second case, we fit the data over the first 400 trials of the generalization tasks (in both cases, the data were also fit in a more conventional fashion—i.e., aggregating over the entire block of trials rather than explicitly modeling performance changes as a function of time—to demonstrate that any outcomes were not specific to the analysis technique).

As can be seen in Figures 4 and S4, when examining just the first 50 trials of data for the generalization task, no significant differences were observed (threshold over first 50 trials: multiple-task trained/untrained comparison—for untrained, 20.12 ± 4.07; for multiple-task trained, 17.52 ± 2.80; $t(16.71) = 0.52$, $p = 0.61$; note that a similar outcome is seen in the first/fourth comparison: for first tasks, −0.14 ± 0.20; for fourth tasks, 0.14 ± 0.33; $t(12) = 0.70$, $p = 0.49$). In other words, this approach would have led to the correct conclusion that no "transfer" was present, but due to the small number of trials involved, it would have resulted in a failure to detect the presence of "learning to learn" (i.e., there would not have been sufficient time for the groups to split apart). Conversely, when examining the first 400 trials as a single block, significant differences were observed for both the trained/untrained and first/fourth comparisons (threshold over first 400 trials: trained/untrained comparison—for untrained, 15.59 ± 1.08; for multiple-task trained, 8.92 ± 0.57; $t(13.9) = 5.47$, $p < 0.001$; first/fourth comparison—for first tasks, 0.42 ± 0.29; for fourth tasks, −0.42 ± 0.19; $t(12) = 2.90$, $p = 0.01$). In essence, by aggregating performance over a large number of trials, one would correctly infer that learning generalization was present but would erroneously conclude that this reflects "transfer" rather than differences in learning rate (i.e., without modeling performance across time, there is no way to differentiate an immediate difference in performance from a rapid splitting apart of performance). This latter issue might be compounded by the use of staircase procedures to estimate thresholds (a common procedure in the field), because the single data point that arises from a staircase procedure is the result of tens, if not hundreds, of trials. Finally, while it is common to utilize some number of practice trials prior to the actual generalization task, this too may result in issues with correctly identifying the form of generalization that is present, as the practice trials could provide an opportunity for two groups to begin splitting apart (in which case, group differences may then be observed as early as the first few trials of the actual generalization task).

We note, though, that although this framework makes a clear prediction that learning multiple tasks with shared task structure should increase learning rate on new tasks that share the same structure, the predictions regarding first-trial performance are much more task dependent. The training tasks in the experiments above were designed in order to minimize the extent to which the shared structure should inform first-trial performance, but this need not be the case. In situations where, for instance, participants bring some knowledge about the new task dimensions, one might expect to see both better initial performance

and faster learning (i.e., if there is a match between both the higher-level structure and some number of the task-level parameters). The types of methodological changes suggested above could thus pave the way toward further addressing what is always the key question for the field—namely, "what" is being learned from training on a given task. The evidence provided here strongly indicates that considering performance on new untrained tasks, both in terms of immediate performance and in terms of learning rate, can serve to differentiate and identify what has been learned via training in a way not provided for by previous methods (e.g., in both "transfer" and "learning to learn," one is applying previous experience to new tasks, but in the former case, that knowledge is directly applicable to the new task and thus benefits appear immediately, while in the latter case, the knowledge can only serve to shape learning of the new task).

The current data may also suggest the need to further explore a number of other previous results, both within and beyond the domain of perceptual training. For instance, there is a great deal of research on the impact that various complex forms of experience have on perceptual and cognitive skills (e.g., cognitive training, "brain training," etc. [34–37]). As has generally been true of the standard perceptual learning literature, work in these fields has focused exclusively on the "transfer" form of generalization without necessarily utilizing methods that could differentiate "transfer" from "learning to learn." It is thus possible that "learning to learn" has, at times, been misidentified as "transfer" (in the case of positive results) and/or that the "learning to learn" form of generalization was not detected (in the case of non-significant generalization results). The same is also potentially true of work on more complex training regimens within the perceptual learning domain, such as the nicely elaborated "rules-based learning" framework [6]. Here, depending on the "rule" that is learned, one might expect *either* immediate "transfer"—as, for instance, would be predicted if the rule were essentially a template for the to-be-identified target—or "learning to learn"—as would be predicted if the rule was, for instance, more broadly about how to best separate targets from noise. And if rules at various levels of abstraction are learned [38–40], this could result in *both* some degree of immediate "transfer" and some degree of "learning to learn" (or, indeed, one could imagine a situation where one must learn some task statistics before knowing which of several possible rules to adopt). Examining these issues further, though, will require the types of methodological designs and statistical analyses highlighted here that can separate these distinct forms of generalization.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, four figures, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.cub.2017.01.046.

AUTHOR CONTRIBUTIONS

Conceptualization, C.S.G.; Methodology, C.S.G., F.K., A.C., and T.E.G.; Investigation, C.S.G., F.K., and T.E.G.; Analysis, A.C. and F.K.; Writing – Original Draft, C.S.G., F.K., A.C., and C.C.; Writing – Review & Editing, C.S.G., F.K., A.C., C.C., and T.E.G.; Funding Acquisition, C.S.G.; Supervision, C.S.G. and F.K.

ACKNOWLEDGMENTS

This work was supported by Office of Naval Research grants N00014-14-1-0512 and N00014-17-1-2049 to C.S.G. and a University of Wisconsin Office of the Vice Chancellor for Research and Graduate Education Research Fall Research Competition award to C.S.G. The research involved the participation of human subjects and it was approved by the University of Wisconsin-Madison Education and Social/Behavioral Sciences Institutional Review Board. All participants provided written informed consent prior to participation.

Received: July 27, 2016
Revised: December 5, 2016
Accepted: January 23, 2017
Published: March 2, 2017

REFERENCES

1. Fiorentini, A., and Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency. Nature *287*, 43–44.
2. Ball, K., and Sekuler, R. (1982). A specific and enduring improvement in visual motion discrimination. Science *218*, 697–698.
3. Fahle, M. (2004). Perceptual learning: a case for early selection. J. Vis. *4*, 879–890.
4. Jeter, P.E., Dosher, B.A., Petrov, A., and Lu, Z.L. (2009). Task precision at transfer determines specificity of perceptual learning. J. Vis. *9*, 1–13.
5. Xiao, L.Q., Zhang, J.Y., Wang, R., Klein, S.A., Levi, D.M., and Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. Curr. Biol. *18*, 1922–1926.
6. Zhang, J.Y., Zhang, G.L., Xiao, L.Q., Klein, S.A., Levi, D.M., and Yu, C. (2010). Rule-based learning explains visual perceptual learning and its specificity and transfer. J. Neurosci. *30*, 12323–12328.
7. Snell, N., Kattner, F., Rokers, B., and Green, C.S. (2015). Orientation transfer in vernier and stereoacuity training. PLoS ONE *10*, e0145770.
8. Green, C.S., Kattner, F., Siegel, M.H., Kersten, D., and Schrater, P.R. (2015). Differences in perceptual learning transfer as a function of training task. J. Vis. *15*, 5.
9. Deveau, J., Ozer, D.J., and Seitz, A.R. (2014). Improved vision and on-field performance in baseball through perceptual learning. Curr. Biol. *24*, R146–R147.
10. Thorndike, E.L., and Woodworth, R.S. (1901). The influence of improvement in one mental function upon the efficiency of other functions. Psychol. Rev. *8*, 247–261.
11. Singley, M.K., and Anderson, J.R. (1989). The Transfer of Cognitive Skill (Harvard University Press).
12. Harlow, H.F. (1949). The formation of learning sets. Psychol. Rev. *56*, 51–65.
13. Thrun, S., and Pratt, L.e. (1998). Learning to Learn (Kluwer Academic Publishers).
14. Kemp, C., Goodman, N.D., and Tenenbaum, J.B. (2010). Learning to learn causal models. Cogn. Sci. *34*, 1185–1243.
15. Bavelier, D., Green, C.S., Pouget, A., and Schrater, P. (2012). Brain plasticity through the life span: learning to learn and action video games. Annu. Rev. Neurosci. *35*, 391–416.
16. Braun, D.A., Mehring, C., and Wolpert, D.M. (2010). Structure learning in action. Behav. Brain Res. *206*, 157–165.
17. Bejjanki, V.R., Zhang, R., Li, R., Pouget, A., Green, C.S., Lu, Z.L., and Bavelier, D. (2014). Action video game play facilitates the development of better perceptual templates. Proc. Natl. Acad. Sci. USA *111*, 16961–16966.
18. Botvinick, M.M. (2008). Hierarchical models of behavior and prefrontal function. Trends Cogn. Sci. *12*, 201–208.
19. Hinton, G.E. (2007). Learning multiple layers of representation. Trends Cogn. Sci. *11*, 428–434.

20. Michel, M.M., and Jacobs, R.A. (2007). Parameter learning but not structure learning: a Bayesian network model of early perceptual learning. J. Vis. *7*, 4.

21. Edelman, S., and Duvdevani-Bar, S. (1997). A model of visual recognition and categorization. Philos. Trans. R. Soc. Lond. B Biol. Sci. *352*, 1191–1202.

22. McLaren, I.P.L. (1997). Categorization and perceptual learning: an analogue of the face inversion effect. Q. J. Exp. Psychol. A *50*, 257–273.

23. Mitchell, C., and Hall, G. (2014). Can theories of animal discrimination explain perceptual learning in humans? Psychol. Bull. *140*, 283–307.

24. Garrigan, P., and Kellman, P.J. (2008). Perceptual learning depends on perceptual constancy. Proc. Natl. Acad. Sci. USA *105*, 2248–2253.

25. Kellman, P.J. (2004). Perceptual learning. In Stevens' Handbook of Experimental Psychology, Third Edition, H. Pashler, ed. (Wiley).

26. Kattner, F., Cox, C.R., and Green, C.S. (2016). Transfer in rule-based category learning depends on the training task. PLoS ONE *11*, e0165260.

27. Sagi, D. (2011). Perceptual learning in vision research. Vision Res. *51*, 1552–1566.

28. Braun, D.A., Aertsen, A., Wolpert, D.M., and Mehring, C. (2009). Motor task variation induces structural learning. Curr. Biol. *19*, 352–357.

29. Dickinson, A., and Pearce, J.M. (1977). Inhibitory interactions between appetitive and aversive stimuli. Psychol. Bull. *84*, 690–711.

30. Griffiths, T.L., and Tenenbaum, J.B. (2005). Structure and strength in causal induction. Cognit. Psychol. *51*, 334–384.

31. Brown, A.L., and Kane, M.J. (1988). Preschool children can learn to transfer: learning to learn and learning from example. Cognit. Psychol. *20*, 493–523.

32. Tenenbaum, J.B., and Griffiths, T.L. (2001). Generalization, similarity, and Bayesian inference. Behav. Brain Sci. *24*, 629–640.

33. Green, C.S., and Bavelier, D. (2012). Learning, attentional control, and action video games. Curr. Biol. *22*, R197–R206.

34. Jaeggi, S.M., Buschkuehl, M., Jonides, J., and Perrig, W.J. (2008). Improving fluid intelligence with training on working memory. Proc. Natl. Acad. Sci. USA *105*, 6829–6833.

35. Jaeggi, S.M., Buschkuehl, M., Jonides, J., and Shah, P. (2011). Short- and long-term benefits of cognitive training. Proc. Natl. Acad. Sci. USA *108*, 10081–10086.

36. Kellman, P.J., Massey, C.M., and Son, J.Y. (2010). Perceptual learning modules in mathematics: enhancing students' pattern recognition, structure extraction, and fluency. Top. Cogn. Sci. *2*, 285–305.

37. Anguera, J.A., Boccanfuso, J., Rintoul, J.L., Al-Hashimi, O., Faraji, F., Janowich, J., Kong, E., Larraburo, Y., Rolle, C., Johnston, E., and Gazzaley, A. (2013). Video game training enhances cognitive control in older adults. Nature *501*, 97–101.

38. Tenenbaum, J.B., Kemp, C., Griffiths, T.L., and Goodman, N.D. (2011). How to grow a mind: statistics, structure, and abstraction. Science *331*, 1279–1285.

39. Kemp, C., Perfors, A., and Tenenbaum, J.B. (2007). Learning overhypotheses with hierarchical Bayesian models. Dev. Sci. *10*, 307–321.

40. Lake, B.M., Salakhutdinov, R., and Tenenbaum, J.B. (2015). Human-level concept learning through probabilistic program induction. Science *350*, 1332–1338.

Appendix 4. Functional form of perceptual learning supplemental information

Texture Oddball Detection Task details

Participants completed 4 blocks of texture oddball detection perceptual training on each of 2 days. Each block included 210 trials. Training lasted between 45 and 60 minutes on each day. Prior to block 1 and block 7 the participant completed 4 practice trials using very easy stimuli at the longest SOA.

On each trial one stimulus texture (black lines on a white background) was centrally presented for an amount of time that varied between trials (i.e., Stimulus Onset Asynchrony; SOA: .03, .06, .09, .12, .24, .36, .48 seconds). Stimuli were a 7 by 7 grid of identical oriented lines, .37 degrees of visual angle in length and with .72 degrees of angle between center points, with one line being at an oddball angle on a random 50% of trials (counterbalanced with SOA). Trial structure is shown in Figure 6. A standard keyboard left arrow key was used to indicate an "all same" response while the right arrow key was used to indicate an "oddball present" response.

Dot-motion Direction Discrimination Task details

Participants completed 4 blocks of dot-motion perceptual training on each of 2 days. Each block included 200 trials. Training lasted between 45 and 60 minutes on each day. Prior to block 1 and block 7 the participant completed 4 practice trials using very easy stimuli and double the presentation time of normal trials.

On each trial two stimuli were presented inside a centrally-located circular aperture with a radius of 4 degrees of visual angle and a medium-grey background (i.e., halfway between minimum and maximum luminance for the screen). Each stimulus was a field of 400 uniformly distributed black dots with diameters of .09 degrees of visual angle. Dot stimuli moved at a continuous speed of 10 degrees of visual angle per second. The first stimulus was presented for 500ms, followed by a 200ms blank aperture, followed by a second stimulus for 500ms. After the second stimulus a blank aperture was presented until the participant response. Correct responses were followed by a high-pitched feedback tone while incorrect responses were followed by a low-pitched feedback tone.

Each trial's two stimulus directions were either the same or different, with each of the two participant response keys being mapped to either "same" or "different" responses. Easy-condition participants' trial stimulus directions were either 0° or 8° apart and were ±4° from a block reference angle. Difficult-condition participants' trial stimulus directions were either 0° or 4° apart and were ±2° from the block reference angle. The first 6 blocks (1200 training trials) utilized a reference angle of 40° while the last 2 blocks (400 generalization trials) utilized a 130° reference angle.

Model code:

Dot-motion direction discrimination

```
m_tef_exp3 <- TEfitAll(motDat[,c('dPrime',"totalTrialNum","isTransfer")],
                  errFun='ols',changeFun='expo', covarTerms=list(pAsym=F),
                  bootPars=list(nBoots=nBoot,bootPercent=.8),
                  control = tef_control(suppressWarnings=T,nTries=nTries,y_lim=c(0,5)),
                  groupingVar=motDat$subID,groupingVarName = 'subID')

 m_tef_exp4 <- TEfitAll(motDat[,c('dPrime',"totalTrialNum","isTransfer")],
                  errFun='ols',changeFun='expo_double', covarTerms=list(pAsym=F),
                  bootPars = list(nBoots=nBoot,bootPercent=.8),
                  control = tef_control(suppressWarnings=T,nTries=nTries,y_lim=c(0,5)),
                  groupingVar=motDat$subID,groupingVarName = 'subID')

 m_tef_pow3 <- TEfitAll(motDat[,c('dPrime',"totalTrialNum","isTransfer")],
                  errFun='ols',changeFun='power', covarTerms=list(pAsym=F),
                  bootPars = list(nBoots=nBoot,bootPercent=.8),
                  control = tef_control(suppressWarnings=T,nTries=nTries,y_lim=c(0,5)),
                  groupingVar=motDat$subID,groupingVarName = 'subID')

 m_tef_pow4 <- TEfitAll(motDat[,c('dPrime',"totalTrialNum","isTransfer")],
                  errFun='ols',changeFun='power4', covarTerms=list(pAsym=F,pPrevTime=F),
                  bootPars = list(nBoots=nBoot,bootPercent=.8),
                  control = tef_control(suppressWarnings=T,nTries=nTries,y_lim=c(0,5)),
                  groupingVar=motDat$subID,groupingVarName = 'subID')

 m_tef_weib <- TEfitAll(motDat[,c('dPrime',"totalTrialNum","isTransfer")],
                  errFun='ols',changeFun='weibull', covarTerms=list(pAsym=F),
                  bootPars = list(nBoots=nBoot,bootPercent=.8),
                  control = tef_control(suppressWarnings=T,nTries=nTries,y_lim=c(0,5),
                        shape_lim=c(-2,2)),
                  groupingVar=motDat$subID,groupingVarName = 'subID')
```

Texture oddball detection

```
 m_tef_exp3 <- TEfitAll(texDat[,c('Corr',"totalTrialNum","SOA","isTransfer")],
                        errFun='bernoulli',linkFun = list(link='weibull',weibullX='SOA'),
                       changeFun='expo', bootPars=list(nBoots=nBoot,bootPercent=.8),
                        covarTerms=list(threshAsym=F),
                       control = tef_control(suppressWarnings=T,nTries=nTries),
                       groupingVar = texDat$subID,groupingVarName = 'subID')

m_tef_exp4 <- TEfitAll(texDat[,c('Corr',"totalTrialNum","SOA","isTransfer")],
                         errFun='bernoulli',linkFun = list(link='weibull',weibullX='SOA'),
                      changeFun='expo_double', bootPars=list(nBoots=nBoot,bootPercent=.8),
                      covarTerms=list(threshAsym=F),
                      control = tef_control(suppressWarnings=T,nTries=nTries),
                      groupingVar = texDat$subID,groupingVarName = 'subID')
m_tef_pow3 <- TEfitAll(texDat[,c('Corr',"totalTrialNum","SOA","isTransfer")],
                         errFun='bernoulli',linkFun = list(link='weibull',weibullX='SOA'),
                      changeFun='power',  bootPars=list(nBoots=nBoot,bootPercent=.8),
                      covarTerms=list(threshAsym=F),
                      control = tef_control(suppressWarnings=T,nTries=nTries),
                      groupingVar = texDat$subID,groupingVarName = 'subID')

m_tef_pow4 <- TEfitAll(texDat[,c('Corr',"totalTrialNum","SOA","isTransfer")],
                          errFun='bernoulli',linkFun = list(link='weibull',weibullX='SOA'),
                          changeFun='power4',  bootPars=list(nBoots=nBoot,bootPercent=.8),
                          covarTerms=list(threshAsym=F),
                          control = tef_control(suppressWarnings=T,nTries=nTries),
                          groupingVar = texDat$subID,groupingVarName = 'subID')

m_tef_weib <- TEfitAll(texDat[,c('Corr',"totalTrialNum","SOA","isTransfer")],
                          errFun='bernoulli',linkFun = list(link='weibull',weibullX='SOA'),
                          changeFun='weibull',  bootPars=list(nBoots=nBoot,bootPercent=.8),
                          covarTerms=list(threshAsym=F),
                          control = tef_control(suppressWarnings=T,nTries=nTries),
                          groupingVar = texDat$subID,groupingVarName = 'subID')
```

Appendix 5. Cochrane et al. (2018)

Rapid Learning in Early Attentional Processing: Bayesian Estimation of Trial-by-Trial Updating

Aaron Cochrane (akcochrane@wisc.edu)
Vanessa Simmering (simmering@wisc.edu)
Joseph L. Austerweil (austerweil@wisc.edu)
C. Shawn Green (cshawn.green@wisc.edu)
Department of Psychology, 1202 W. Johnson Street
Madison, WI 53706 USA

Abstract

All agents must constantly learn from dynamic environments to optimize their behaviors. For instance, it is necessary in new environments to learn how to distribute attention – i.e., which stimuli are relevant, and thus should be selected for greater processing, and which are irrelevant, and should be suppressed. Despite this, many experiments implicitly assume that attentional control is a static process (by averaging performance over large blocks of trials). By developing and utilizing new statistical tools, here we demonstrate that the effect of flanking items on response times to a central item (often utilized as an index of attentional control) is systematically and continuously influenced through time by the statistics of the flanking items. We discuss the implications of this finding from the perspective of examining individual differences – where traditional data analysis approaches may confound the rate at which attentional filtering changes through time with the asymptotic ability to filter.

Keywords: Learning; attention; statistical inference; Bayesian analysis

Introduction

In human, animal, and artificial cognitive architectures, learning to utilize available information for goal-directed behavior is a crucial ability. Critically, in nearly all theoretical models, learning is viewed as an inherently continuous process – with each new data point that is sampled resulting in some concomitant change in knowledge and behavior. One consequence of this is that, even in cases where huge amounts of data have already been sampled (and thus each new data point will change behavior by only a small amount), there remain very few behaviors that would be posited to be fully static or unchanging over time.

Despite this theoretical foundation, in practice the standard analysis approach taken for tasks in the psychological literature implicitly assumes that behavior is in fact static over some period of time (if not the entirety of the task experience). This tendency is even seen in the study of learning – where it has traditionally been quite common to examine performance divided into arbitrary discrete timescales. That is, performance is typically divided into methodologically useful units, such as blocks of training or testing. Data within the block is analyzed under the assumption that the same process generated the data within the entirety of a given block. This generally takes the form of aggregating within-block performance using some function or algorithm in order to summarize performance, e.g., percent correct (Ahissar & Hochstein, 1997), logistic psychometric function (Schütt, Harmeling, Macke, & Wichmann, 2016), or Drift Diffusion Model (DDM) parameters (White, Brown, & Ratcliff, 2012; Wiecki, Sofer, & Frank, 2013). While such aggregation has some methodological and analytic utility, in terms of its simplification of the data and behavior, it imposes artificial structure upon learning processes that are theoretically independent of that structure.

Not only is it the case that the effects of interest are almost certainly independent of block structures, but the learning that occurs within blocks may itself be theoretically informative. We have previously shown, by developing and employing a time-continuous data analytic approach for assessing visual perceptual learning task performance, that it is possible to differentiate between two distinct forms of learning generalization. The first type of generalization leads to immediate benefits (i.e., is present from the very first trial of a new task), while the second involves no immediate changes in performance on new tasks, but instead new tasks are learned more quickly. These distinct patterns have enormous theoretical importance, as they are generated via completely different mechanisms. Yet they are impossible to differentiate via traditional data analytic techniques that aggregate performance over large blocks of trials (Kattner, Cochrane, Cox, Gorman, & Green, 2017).

Here we extend the general approach to modeling performance as a continuous-function of time to an area where the potential for learning effects are much more rarely considered – the study of attentional control over peripheral (i.e., non-target) processing. Indeed, the analytic techniques utilized in this domain nearly always implicitly assume that performance is static through time. For instance, such aggregation-based analyses are commonly utilized as individual difference metrics, to identify atypical populations (e.g., ADHD; Westerberg, Hirvikoski, Forssberg, & Klingberg, 2004), to characterize development (Rueda et al., 2004), or to simply benchmark difficulties of a test (Edwards et al., 2006). This is despite the fact that it is unlikely to be the case that participants can enter a task with perfect knowledge regarding the spatial and temporal properties of the task-relevant (i.e., to-be-attended target) stimuli or the spatial and temporal properties of distractors (i.e., the to-be-

ignored stimuli). Here we examine the extent to which learning can be identified and modeled in one extremely common index of attentional control – flanker task performance.

Previous work

Within the study of attentional control, certain domains have largely been understood as automatic and independent of the associations or statistics of the environment and thus reasonably impervious to learning (Treisman, 1985; Wolfe, 1994). However, it has also been recognized that in order for a person to interact optimally with their environment, they must constantly weight the utility of the information available to them at all levels of processing. Indeed, dynamic allocation of attention is a core aspect of human ability to interact with the world. Flexibly adapting attention to the changing demands of the environment allows efficient and accurate goal-directed processing of the relevant information available.

When searching for a target in the visual world, distracting items become increasingly easy to suppress as they become increasingly distinct from the target. In the opposite case, when responses between searched-for items and irrelevant items are opposing, a marked increase in response times (i.e., increase in effort needed) to the relevant items is observed. Remarkably, this occurs even when participants are given explicit instructions regarding where and when the relevant item will occur (as well as any irrelevant distracting items). One paradigm in this vein is the arrow flanker task, derived from Eriksen and Eriksen (1974). In this task participants simply press the right keyboard arrow when a central stimulus is a right-pointing arrow, and they press the left keyboard arrow when the central stimulus is a left-pointing arrow. Two other arrows appear on either side of the central arrow pointing in either the same or opposite direction of the central arrow. When the flanking arrows point in the same direction as the central arrow, response times tend to be faster and more accurate than when flankers point in the opposite direction as the central arrow. The differences between congruent-flanker response times and incongruent flanker-response times are largely understood as slowing that occurs due to processing of response-incompatible stimuli, and the magnitude of this difference is often referred to as the "flanker effect."

The flanker effect has been explored in many settings and interpreted in a wide variety of ways. These primarily involve appeals to a neuropsychological executive function or conflict-resolution mechanism (Fan, McCandliss, Sommer, Raz, & Posner, 2002; Machizawa & Driver, 2011). Through this lens the flanker effect has been correlated with such measures as age (Rueda et al., 2004) and cortical thickness (Westlye, Grydeland, Walhovd, & Fjell, 2011). In each of these paradigms, the flanker effect is interpreted as a stable ability within individuals; in effect, it is seen as a robust index of one's ability to control attention and rapidly suppress distracting information. Unfortunately, this perspective disregards another central aspect of humans' interaction with their environments: The necessity of learning how to weight information appropriately given past experience. While previous research has assumed that psychology tasks (e.g., flanker task) index a constant ability level, we instead posit that learning occurs to some extent (Lehle & Hübner, 2008). This learning in peripheral attention occurs despite the fact that participants are given explicit verbal instructions regarding the time, place, and attributes of the to-be-attended information.

We first demonstrate experimentally, using biased task statistics, that block-level analyses show learning in adult humans' performance on a flanker task. Next, we propose a novel analysis of flanker task response time in which performance is modeled as a function of experience (i.e., trial number). We note that participants were not informed of any learning component to the study. Our analyses show that, even this context, decomposition of performance into parameters of continuous learning reveals the dynamics of humans' interactions with their environments.

Method

Participants and procedure

Forty-seven undergraduate participants from the University of Wisconsin-Madison completed all tasks for course credit. One participant was excluded for missing data. The entire study consisted of three tasks – a flanker task, a Useful Field of View task (UFOV; Ball & Owsley, 1993), and a Multiple Object Tracking task (MOT; Pylyshyn & Storm, 1988). Here, for brevity, we will only consider performance on the flanker task.

The flanker task was modeled after that utilized by Rueda et al (2004). Stimuli were colored fish with arrows overlaid on top pointing in either a leftward or rightward direction. The full flanker task was divided into 5 blocks. Each flanker block included feedback regarding response time and incorrectness. Participants first completed a block of 50 no-flanker trials, second a 250-trial block of either 20%-congruent or 80%-congruent flanker trials (randomly-chosen, with the remaining trials being incongruent), third a 250-trial block of 50%-congruent flanker trials, fourth the biased block that they did not already complete (i.e., 250 trials of either 20%-congruent or 80%-congruent), and fifth a 50-trial block with no flankers. In all cases the participants' task was the same – to indicate the direction of the center fish/arrow as quickly and accurately as possible. Our key questions were whether we would see: (1) differences in performance at a broad scale – in terms of different patterns of response times to the congruent and incongruent trials in the different blocks; and (2) at a continuous time-scale – indicating how such shifts are learned through time.

We note that the other two tasks (UFOV and MOT) were completed between the biased blocks of flanker tasks in order

to reduce monotony and obscure the biasing of the task statistics. For example, by including MOT after the 50% congruent condition and before the 80% congruent condition, we intended that participants would have less carry-over of learning from the 50% congruent condition to the 80% congruent condition.

Analysis

We conducted two Bayesian analyses. The first involved fitting a hierarchical linear model to each block's data. This was designed to test whether, in environments with different statistics, people alter their processing of non-relevant information. The second fit time-evolving weighting parameters to the flanker effect for each plot. This model was able to distinguish which component(s) of learning differ between conditions, as well as demonstrating a novel estimate of continuously changing attentional allocation.

By-block analysis

As an initial demonstration that participants learn to alter their attention in response to changing environmental statistics, a Bayesian multilevel linear model was fit. This model tested the effects of condition compatibility proportion, trial flanker compatibility, and the interaction between these two variables. Block k's free parameters for subject *s*, $\beta^{(s)}_{0k}$, …, $\beta^{(s)}_{3k}$ was drawn from a participant-level distribution, which in turn was drawn from a parent distribution shared by all participants.

$$\tau_A \sim G(.001, .001)$$
$$\beta_i \sim N(0, 100)$$
$$\tau^{(s)}_i \sim G(.001, .001)$$
$$\beta^{(1)}_i, \ldots, \beta^{(S)}_i \mid \beta_i \sim N(\beta_i, \tau_A)$$
$$\beta^{(s)}_{i1}, \ldots, \beta^{(s)}_{ik} \mid \beta^{(s)}_i \sim N(\beta^{(s)}_i, \tau^{(s)}_i)$$

$$\log RT \sim N(\beta^{(s)}_{0k} + \beta^{(s)}_{1k} * congruence + \beta^{(s)}_{2k} * percent\_congruent + \beta^{(s)}_{3k} * congruence * percent\_congruent, \tau_A)$$

where τ_A is a precision parameter for the data distribution and $\tau^{(s)}_i$ is a precision parameter shared across blocks per subject.

This model considered only the last 200 trials in each block in order to characterize asymptotic performance. The predicted outcome of this model was that response times to congruent and incongruent trials would be different from one another (as has been seen in all previous research, with congruent RTs being faster than incongruent RTs), but with these differences themselves differing across varying levels of task statistics, meaning that participants had in fact shifted their behavior based upon the task statistics (i.e., had learned). This response time difference should be evident when controlling for trial congruence as well as individual differences in overall response times. This result would provide evidence that the following analysis, on the time course of learning, would be justified.

By-trial analysis

After testing for block-wise differences between conditions in the magnitude of the flanker effect, we defined a generative process that we hypothesized would give rise to continuous changes in the flanker effect. Fitting parameter estimates to this process would provide hierarchical estimates of the inter-individual and intra-individual variations in the adaptation of attention to environmental statistics.

This analysis assumed two interacting processes. First, that each individual has a stable, domain-general speed-of-processing (SoP) ability that indexes how fast that person can perceive, attend to, and react to their environments (Conway, Cowan, Bunting, Therriault, & Minkoff, 2002). In the flanker task, this would be akin to the response time to the central stimulus when disregarding any effect of the peripheral stimuli. Second, there is a flanker-congruence related offset to the baseline response time. Here that offset is modeled as an additive shift to the baseline on a log scale, which translates to a multiplicative shift in raw response times. Approaches to flanker analysis that equate congruent-flanker trials with no-flanker trials would parameterize this relation as simply an additive component to the baseline. However, in order to remain sensitive to the possibility of the shift adding to the response time in the incongruent-flanker condition while subtracting from the response time in the congruent-flanker condition (i.e., speeding), here the shift is parameterized as symmetrically adding or subtracting to a central baseline log SoP response time ability. That is, we maintain the possibility that participants use congruent flankers to speed up their response times (noting though that the high-level pattern of results with respect to learning should not be strongly dependent on this choice).

Typical individual-differences flanker analyses utilized in the field assume that the congruency-related shift (whether solely positive or not) is stable across the course of the flanker task. Indeed, in order to remain valid, the shift must even be constant across several repetitions the task by a single person. Here that assumption is relaxed. Rather than assume a constant additive shift due to flanker type, the additive shift is assumed to be learned. That is, participants update their attention to flanking items (i.e., their additive shift) throughout the task in response to the utility of attending to peripheral items. Here the additive shift is parameterized as exponential decay as a function of trial number, $(a+b*c^{-t})$. Exponential learning functions are extremely common in many fields, and provide concise characterizations of the time course of learning (Heathcote, Brown, & Mewhort, 2000).

Thus, the generative process assumes the response time on a given trial is the following:

$$\beta_A \sim N(-1,.1) \qquad \tau_A \sim G(100, .1)$$
$$\beta_{asym} \sim N(.1, .01) \qquad \beta_{scale} \sim N(.01, .1)$$
$$\beta_{rate} \sim N(1.1,.1)$$

$$\beta^{(1)}_{SoP}, \ldots, \beta^{(s)}_{SoP} \sim | \beta_{SoP} \; N(\beta_A, \tau_A)$$
$$\beta^{(1)}_{asym}, \ldots, \beta^{(s)}_{asym} | \beta_{asym} \sim N(\beta_{asym}, \tau_A)T(0,2)$$
$$\beta^{(1)}_{scale}, \ldots, \beta^{(s)}_{scale} | \beta_{scale} \sim N(\beta_{scale}, \tau_A)\; T(0,2)$$
$$\beta^{(1)}_{rate}, \ldots, \beta^{(s)}_{rate} | \beta_{rate} \sim N(\beta_{rate}, \tau_A)\; T(-2,2)$$

$$\beta^{(s)}_{asym(1)}, \ldots, \beta^{(s)}_{asym\,(k)} | \beta^{(s)}_{asym} \sim N(\beta^{(s)}_{asym}, \tau_A)$$
$$\beta^{(s)}_{scale(1)}, \ldots, \beta^{(s)}_{scale(k)} | \beta^{(s)}_{scale} \sim N(\beta^{(s)}_{scale}, \tau_A)$$
$$\beta^{(s)}_{rate(1)}, \ldots, \beta^{(s)}_{rate\,(k)} | \beta^{(s)}_{rate} \sim N(\beta^{(s)}_{rate}, \tau_A)$$

$$log\ RT \sim N(\beta^{(s)}_{SoP} + congruence * (\beta^{(s)}_{asym(k)} + \beta^{(s)}_{scale(k)} * \beta^{(s)}_{rate(k)}{}^{-trial}),\ \tau_A)$$

where T(a,b) truncates a distribution to the range (a,b).

Each of the three learning parameters of interest (flanker offset asymptote, scaling and exponent terms) were estimated as normal distributions for each block, with the mean of this normal being drawn from participant-level asymptote (truncated at 0 and 2), scaling (truncated at -2 and 2), and exponent (rate; truncated at 0 and 2) normal distributions. Truncations were imposed at values beyond which model behavior would be qualitatively very different than the theoretical generative model. In particular, the entire peripheral attention term should evaluate to less than 1 in every instance in order to be a sensible fit to the data-generation process (i.e., the difference between incongruent-flanker trials and congruent-flanker trials is never more than 2 in log-RT space).

All other prior distributions were non-truncated, with normal priors for all mu distributions and gamma priors for all gamma distributions and precision distributions. Given the primary interest in comparing between-block within-subjects variation as a function of block statistics, all variation of interest should be caused by the data and not by prior specification.

Results

Bayesian analysis using JAGS implemented in R (Plummer, 2003) was used for parameter approximation. Four chains were burned in for 20,000 samples, then 200,000 samples were drawn for further analysis.

All response time measures were first trimmed to exclude values above 2 seconds and below .05 seconds (120 trials total rejected), as response times outside these bounds are clearly not arising from the processes of interest in this study. In addition, all trials with incorrect responses were excluded (8.9%); further analysis of this incorrect-trial data may be relevant to the core questions of this study, but analysis of this variable was outside the scope of the current paper (see Limitations section below). After this trimming, the remaining 25,064 response times were log-transformed to better approximate normality. Given this, log-transformed response times varied from -2.99 to 0.66 (m= -.99, sd= .237).

Convergence

Bayesian analysis appeared to converge in both models. Visual inspection of trace plots, autocorrelation plots, and Gelman-Rubin plots indicated convergence for the majority of estimated parameters. Five parameters in the by-trial analysis, all of which were block-level estimates of the rate parameter (i.e., $\beta^{(s)}_{rate\,(k)}$), presented clearly problematic traces and autocorrelations. We excluded the five participants with problematic rate parameter traces from the following by-trial analyses, leaving data from 41 participants.

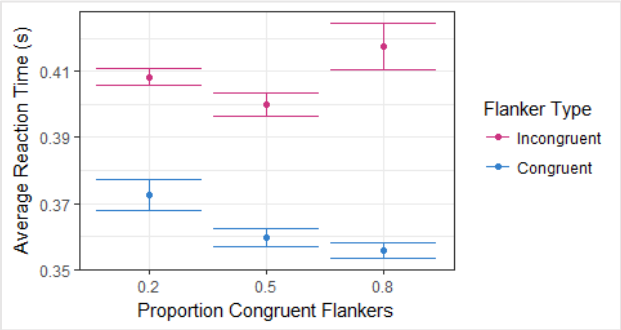


Figure 1. Mean response times, separated by flanker congruence and proportion congruent flankers. Error bars denote 95% confidence intervals across all trials for all participants.

By-block fits

Point estimates of the parameters of interest (i.e., the means of the level-one regression predictors) each provide support for the hypothesis that attention changes with environmental statistics. The 95% credible intervals of these parameters follow the same pattern. This is the case for the effects of flanker congruence (mean= -.081, lower= -.100 , upper = -.062), proportion congruence (mean= -.052, lower= -.114, upper = .009), and the interaction between the two (mean= -.046, lower= -.091 , upper = -.002). For each of these parameters, the vast majority of the mass of the distribution is on one side of zero. It is evident in Figure 1 that this pattern supports the hypothesized effects; in situations where the participant sees mostly facilitative non-targets, they are faster in responding to congruent-flanker trials while also being slower on incongruent-flanker trials in these situations. In essence, when most of the flankers are congruent, there is an advantage in reducing the extent to which these flankers are filtered. This produces faster RTs on congruent trials, but then causes disproportionate slowing on incongruent trials. Meanwhile, in conditions where most of the trials are incongruent, there is virtue in strongly filtering all flankers. Note a lesser facilitatory effect is then seen for the congruent

trials, but the magnitude of the drop-off in RT on incongruent trials is reduced.

Although we explore this in greater detail below, the data already indicates that such behavioral shifts must be learned over many trials. For instance, when subsets of the 50-50 condition are analyzed – either short runs in which four congruent trials were followed by an incongruent trial, or four incongruent trials were followed by a congruent trial (effectively creating miniature "80-20" or "20-80" conditions) the change in response time from the fourth to the fifth trials is no different than the overall differences between congruent and incongruent trials in the 50-50 condition (both $|t|<1.1$). This suggests that the large-scale differences seen across the different blocks are the result of a longer-term learning process. The following by-trial analysis further tested the time course of learning.

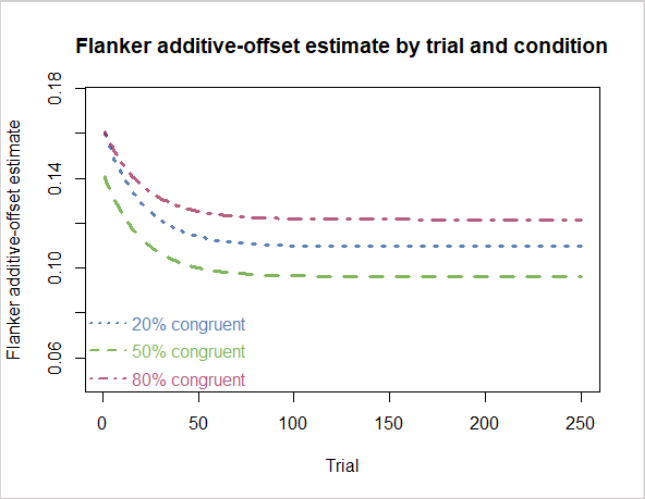


Figure 2. Mean by-trial half-flanker effect (i.e., the additive component of the flanker effect, which would have a subtractive component mirrored across y=0). Separated by block type (percent of congruent trials).

By-trial fits

Means of evaluated fits are shown in Figure 2. The three main parameters were asymptote, which indicated the flanker effect size after learning, scale, which indicated the magnitude and direction of learning, and rate, which indicated the relative speed of learning. We conducted preliminary frequentist nonparametric comparisons of fit parameters in the extreme conditions via within-subjects Wilcoxon signed-rank tests that compared the 20% congruent to the 80% congruent conditions. The difference between the rate parameters in 20%-congruent condition (median = 1.0517) and 80%-congruent condition (median = 1.0502) was 0.0015, V = 536, p = 0.176. The difference between the asymptote parameters in 20%-congruent condition (median = 0.109) and 80%-congruent condition (median = 0.129) was 0.019, V = 301, p = 0.095. The difference between the scale parameters in 20%-congruent condition (median = 0.051) and 80%-congruent condition (median = 0.041) was 0.010, V = 771, p<.001. These analyses provide preliminary evidence for reliable differences in learning scale, but not the other parameters, between conditions. Scale determines magnitude as well as direction of learning, making it a reasonable parameter to expect to differ between conditions if we believe that people truly are learning to behave differently.

We next tested three frequentist multilevel models using R packages *lme4* and *pbkrtest*, one for linear changes in each of the three exponential parameters due to changes in congruency statistics.

lmer(param~propCongruent+(propCongruent|subject))

Each model used the proportion of congruent trials in a block to predict the fit parameter value, while controlling for participant-level random effects. The rate parameter was not reliably predicted by task statistics, $F(1,44.7)=.817$, $p>.35$. The asymptote likewise fell short of conventional statistical significance, $F(1,43.1)=3.97$, $p=.053$. In contrast, but in concurrence with the Wilcoxon test reported above, the scale parameter was linearly predicted by varying flanker-congruence proportions, $F(1,44.0)=26.91$, $p<.001$.

Limitations

These analyses have certain weaknesses and shortcomings. For example, the apparent nonmonotonicity of fits with regard to task statistics could be an artifact of block order effects that indicates learning to learn. In addition, this work, meant as a preliminary demonstration, utilizes a simplistic measure of flanker performance. Response times for a given trial are assumed to be additively shifted from a baseline (in a log-transformed scale), while trials with incorrect responses are omitted. Further work should explore continuous learning-related changes via models that capture both response time and accuracy (as in Drift Diffusion Models - DDMs), as these are more meaningful decompositions of performance than the only-correct log-transformed RTs reported here. Doing so may require longer learning blocks, as DDMs with relatively high numbers of parameters are unlikely to recover reliable estimates of learning parameters given a mere 250 learning trials per condition. While hierarchical modelling would somewhat alleviate these concerns by providing stability (i.e., lower-level parameters could only be estimated from the distributions of higher-level parameters), we refrained from testing these models with high numbers of free parameters here. One direction for future work could be to apply hierarchical DDM parameters with covariates (Wiecki et al., 2013) to the problem of trial-by-trial learning by specifying a functional form (e.g., exponential decay). Many other additional parameters could be fit as well, such as changing SoP values or asymmetric flanker effects (e.g., additive effects due to incongruent trials being larger than subtractive effects due to congruent trials).

Discussion

Here we demonstrate that bottom-up attention is reliably influenced by environmental statistics. That is, the degree of filtering demonstrated by a participant is a function of some amount of learning – rather than fully reflecting a static ability. We provide evidence for a quantitative dissociation between the dynamics of learning and the stable individual differences that interact to give rise to the overall pattern of behavior in the flanker task. The scale (indicating size and/or direction) of learning is clearly changed by environmental statistics. The asymptote appears to be changed as well, although our data indicates that this change is not linear or even monotonic (see Figure 2). While many questions remain to be examined regarding learning in attentional tasks, this first step provides impetus to further address how to best quantitatively decompose behavior in single tasks into separate processes, including a learning process.

The key implication of this work is thus that individual differences approaches to attention, and cognition more generally, would benefit from integrating analyses sensitive to the effects of learning. Performance differences between individuals due to learning may be mechanistically distinct from individual differences arising from, for example, stable differences in distractor suppression. Furthermore, such considerations may shed light on other areas of the field – for instance, in examining test-retest reliabilities. Indeed, the test-retest reliability of flanker tasks has typically not been found to be high. However, this may be due to failing to account for learning from test to test. By implementing experimental paradigms and analytical methods capable of identifying the relative contributions of these processes, further light may be shed on the mechanistic underpinnings of a wide array of typical processing (e.g., fluid intelligence) as well as atypical (e.g., ADHD, anxiety).

References

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*(6631), 401–406.

Ball, K., & Owsley, C. (1993). The useful field of view test: a new technique for evaluating age-related declines in visual function. *Journal of the American Optometric Association*, *64*(1), 71–79.

Conway, A. R. ., Cowan, N., Bunting, M. F., Therriault, D. J., & Minkoff, S. R. . (2002). A latent variable analysis of working memory capacity, short-term memory capacity, processing speed, and general fluid intelligence. *Intelligence*, *30*(2), 163–183.

Edwards, J., Ross, L., Wadley, V., Clay, O., Crowe, M., Roenker, D., & Ball, K. (2006). The useful field of view test: Normative data for older adults. *Archives of Clinical Neuropsychology*, *21*(4), 275–286.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*(3), 340–347.

Heathcote, A., Brown, S., & Mewhort, D. J. (2000). The power law repealed: the case for an exponential law of practice. *Psychonomic Bulletin & Review*, *7*(2), 185–207.

Kattner, F., Cochrane, A., Cox, C. R., Gorman, T. E., & Green, C. S. (2017). Perceptual Learning Generalization from Sequential Perceptual Training as a Change in Learning Rate. *Current Biology*, *27*(6), 840–846.

Lehle, C., & Hübner, R. (2008). On-the-fly adaptation of selectivity in the flanker task. *Psychonomic Bulletin & Review*, *15*(4), 814–818.

Machizawa, M. G., & Driver, J. (2011). Principal component analysis of behavioural individual differences suggests that particular aspects of visual working memory may relate to specific aspects of attention. *Neuropsychologia*, *49*(6), 1518–1526.

Plummer, M. (2003). JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. *3rd International Workshop on Distributed Statistical Computing*, *124*.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197.

Rueda, M. R., Fan, J., McCandliss, B. D., Halparin, J. D., Gruber, D. B., Lercari, L. P., & Posner, M. I. (2004). Development of attentional networks in childhood. *Neuropsychologia*, *42*(8), 1029–1040.

Schütt, H. H., Harmeling, S., Macke, J. H., & Wichmann, F. A. (2016). Painfree and accurate Bayesian estimation of psychometric functions for (potentially) overdispersed data. *Vision Research*, *122*, 105–123.

Treisman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics, and Image Processing*, *31*, 156–177.

Westerberg, H., Hirvikoski, T., Forssberg, H., & Klingberg, T. (2004). Visuo-spatial working memory span: a sensitive measure of cognitive deficits in children with ADHD. *Child Neuropsychology*, *10*(3), 155–161.

Westlye, L. T., Grydeland, H., Walhovd, K. B., & Fjell, A. M. (2011). Associations between Regional Cortical Thickness and Attentional Networks as Measured by the Attention Network Test. *Cerebral Cortex*, *21*(2), 345–356.

White, C. N., Brown, S., & Ratcliff, R. (2012). A test of Bayesian observer models of processing in the Eriksen flanker task. *Journal of Experimental Psychology. Human Perception and Performance*, *38*(2), 489–497.

Wiecki, T. V., Sofer, I., & Frank, M. J. (2013). HDDM: Hierarchical Bayesian estimation of the Drift-Diffusion Model in Python. *Frontiers in Neuroinformatics*, *7*.

Wolfe, J. M. (1994). Guided search 2.0: A revised model of visual search. *Psychonomic Bulletin & Review*, *1*, 202–238.

Appendix 6. Cochrane, Simmering & Green (under review)

Modulation of compatibility effects in response to experience: Two tests of initial and sequential learning

Aaron Cochrane*[1], Vanessa Simmering[1,2], C. Shawn Green[1]

[1]Department of Psychology, University of Wisconsin – Madison, Madison, WI, USA

[2]ACTNext at ACT, Inc, Iowa City, Iowa, USA

*corresponding author: 1202 W Johnson St, Madison, WI 53706; (509)366-1379; akcochrane@wisc.edu

Abstract

Attentional control is a key component of goal-directed behavior. Modulation of this control in response to the statistics of the environment allows for flexible processing or suppression of relevant and irrelevant items in the environment. Modulation robustly occurs in compatibility-based attentional tasks, where incompatibility-related slowing is reduced when incompatible events are likely (i.e., the Proportion Compatibility Effect; PCE). The PCE implicates dynamic changes in the measured compatibility effects that are central to fields of study such as attention, executive functions, and cognitive control. In these fields, stability in compatibility effects are generally assumed, which may be problematic if individual or group differences in measured compatibility effects may arise from differences in statistical learning speed or magnitude. Further, the sequential nature of many studies may lead the learning of certain statistics to be inadvertently applied to future behaviors. Here we report tests of learning the PCE across conditions of task statistics and sequential blocks. We then test for the influence of feedback on the development of the PCE. We find clear evidence for the PCE but no conclusive evidence for task learning. Instead, in the absence of feedback a decrement of performance was observed from the beginning to the end of participant's first blocks. Initial experience with more incompatible trials selectively mitigated performance decreases in a subsequent block. Despite the lack of behavioral changes associated with patterns of learning, systematic within-task changes in compatibility effects remain an important possible source of variation in a wide range of attention research.

Introduction

In a seminal paper, Ericksen and Ericksen (1974) demonstrated the flexibility and ubiquity of response competition as a window into visual attention. Critically, the tasks that were developed in this vein of research frequently necessitated that participants learn arbitrary stimulus-response mappings (e.g., a *S* stimulus associated with a right index finger button press). Accordingly, the associated response competition effects that were of primary importance as dependent variables in these tasks were likewise necessarily the products of learning. Indeed, participants did not enter the tasks already having reaction times that were slower when a particular button needed to be pressed in the context of a particular stimulus, as compared to when the same button needed to be pressed in the context of some other stimulus. Instead these patterns had to emerge as the participants internalized the relationship between the various stimuli and button presses (i.e., response competition).

This perspective aligns with the overarching idea that the behavioral markers of attentional phenomena often require certain types of experience in order to be manifested. This is true, for instance, of the *Proportion Compatibility Effect* (PCE), which is observed when participants experience unequal numbers of trials that include compatible (response-congruent) or incompatible (response-incongruent) distractors. More specifically, in the PCE, experimental conditions that involve more frequent response competition (i.e., a greater amount or degree of incompatible responses between targets and distractors) are associated with a reduction in response time (RT) differences due to compatibility effects (Braem et al., 2019; Gratton et al., 1992; Logan & Zbrodoff, 1979).

Several potential sources of this pattern have been proposed. One possibility is that repeated experiences with response conflict causes an augmentation of attentional control (Bugg & Crump, 2012; Gratton et al., 1992; Lehle & Hübner, 2008). This adjustment would decrease the influence of interfering distractors and in turn produce a convergence between compatible and incompatible trials. In the limit, if there was a perfect attentional filtering process, and thus no processing of interfering distractors, RTs should be the same regardless of whether the distractors were response-compatible or response-incompatible (i.e., it would be as if the distractors were “not there”). A second set of proposals has characterized the PCE as arising from an interaction between task difficulty and low-level learning processes (Abrahamse et al., 2016; Schmidt, 2016). Broadly, according to one version of this perspective, more-frequent experiences (i.e., incompatible or compatible trials) are learned most and thus experience a disproportionate decrease in RT. However, the underlying difficulty of incompatible trials nevertheless causes RTs for these trials to remain higher than compatible trials. This theory therefore also predicts a

convergence in compatible and incompatible RT when incompatible trials are frequent. Finally, a third possible explanation for the PCE notes that while experience with multiple trials is clearly necessary to observe the PCE, this experience does not necessarily need to be on a timescale over which the statistics of the task could be adapted to or learned. Instead, short-range trial-to-trial interactions could cause brief variations in behavior that compound to demonstrate the PCE (e.g., the *Congruency Sequence Effect;* Gratton et al., 1992). Importantly though, short-term effects such as priming are not exclusive of longer-term learning or cognitive control modulation in response to task statistics, and in practice the mechanisms are likely to interact (Davelaar & Stevens, 2009).

While the models above were primarily developed to capture PCE effects that emerged within a short time-scale (e.g., within a block of trials or even within pairs of trials), another implication of a learning-centered perspective on the PCE is the possibility of even longer-term dependencies. One such possibility is that earlier experience with an attentionally-demanding task may implicitly teach certain patterns of attentional modulation, rapid inference regarding stimuli, expectations regarding motor responses, or other context-bound information relevant to successful task performance (Schmidt & Weissman, 2014). Longer-term dependencies may not be constrained to just stimulus-response mappings or short-term attentional modulations; future learning may itself be altered by the previous learning environment (Kattner et al., 2017). That is, initial task experience may implicitly teach participants something about the global task context which has longer-term effects on response times for compatible or incompatible trials.

Consistent with these ideas, block-to-block carry-over effects have been observed in PCE. For example, using a Stroop task, Abrahamse et al. (2013) found that majority-compatible blocks were associated with much larger compatibility effects when they were the first block experienced than when preceded by a majority-incompatible block. The magnitude of compatibility effects in majority-incompatible blocks meanwhile was similar regardless of prior experience. Abrahamse (2013) attributed these asymmetrical shifts between low-to-high vs high-to-low proportions of compatible experiences to attentional modulation that unfolded over the course of hundreds of trials. Specifically, when participants' first experience was with a majority-incompatible block, this resulted in a change of attentional focus that not only reduced the magnitude of the compatibility effect within that block (i.e., as would be seen in the typical PCE), but that persisted into the next block (thereby reducing the magnitude of the compatibility effect in that next block as well). Critically, under certain conditions, it was posited that performance shifted too slowly to be captured within their experimental timescale (i.e., the 240 or 288 trials that were

utilized). Furthermore, averaging across all RTs within experimental blocks precluded inferences regarding within-block change.

Importantly, the need to consider the possibility of time-evolving processes in tasks where response competition effects are the primary dependent variables of interest reaches far beyond theoretical questions regarding statistical learning or adaptation of attention. Response competition measures have been utilized in studies ranging from individual differences in cortical anatomy (Westlye et al., 2011) to cognitive training (Rueda et al., 2005) and the effects of psychoactive substances (Bailey et al., 2016), and similar measures are central to influential theories of attentional networks (Fan et al., 2002; Petersen & Posner, 2012; Posner & Petersen, 1990) and executive functions (Diamond, 2013; Miyake, 2000). In each of these domains, aggregate measures of response competition have been used as an index of the effectiveness of low-level control and selection processes, yet the extent to which these processes are being contextually adapted is rarely explicitly examined in these settings.

There are several possible, but not well-examined, implications of the PCE for these broader research domains that make use of response competition measures. As noted above, the PCE is a context-dependent change in response competition measures (i.e., the PCE must necessarily unfold over time as an interaction between person and environment because differences in the number of compatible versus incompatible trials only emerge through time). It is unclear though the extent to which between-participant variation in response competition, which is of interest to many areas of cognitive psychology, arises due to between-participant differences in (1) stable trait-like abilities or (2) magnitude or rate of adaptation to a given context. Further, attentional modulations and statistical learning may influence participants on timescales reaching beyond single blocks of trials, leading to complex sequential effects in the measurement of response competition. Specifically, cognitive theories and empirical results interpreting flanker-task response competition as a static quantity may instead be inadvertently observing and interpreting varying rates of change through time.

One complication in linking the broader literature that has made use of response competition measures to PCE research is that one core aspect of the task design – the presence or absence of feedback – tends to differ across these domains. Indeed, many research domains have employed task versions that do not include explicit feedback (B. A. Eriksen & Eriksen, 1974; Fan et al., 2002; Miyake, 2000). In many cases, whether implicitly or explicitly, this methodological decision may arise from the assumption that behavior is more stable in the absence of additional signals from the environment (i.e., it could reduce the extent to which participants learn via simple experience with the task itself). In contrast, PCE effects

have most commonly been observed in situations where participants are provided with informative feedback (i.e., regarding response time, accuracy, or both; Gratton et al., 1992; Schmidt & Weissman, 2014; Wenke et al., 2015). Even when comparing explicit instruction-based learning to lower-level statistical learning, Wenke and colleagues (2015) included feedback in all experimental conditions.

Here we examined several issues regarding the possible time-evolving nature of flanker effects and the PCE. Because arrow-based flanker tasks are very common in the broader literature on attentional control (Bailey et al., 2016; Davelaar & Stevens, 2009; Fan et al., 2002; Rueda et al., 2005; Sidarus et al., 2019), while PCE effects have typically been primarily investigated in Stroop-like or Simon-like tasks, we first further confirmed that the canonical PCE was observed in participants' initial block of an arrow-flanker task (with different participants receiving 20%, 50%, or 80% of compatible trials in this first block). Next, by modeling RT change as a function of time, we examined whether we could find explicit evidence of the PCE emerging within this first block of trials. Then, by having participants complete a second block of the flanker task with a different proportion of compatible trials than they had experienced during their first block, we assessed whether, after controlling for the second block's proportion compatible, the change in RTs evident in the second block is systematically related to the first block's proportion compatible (i.e., whether there is carry-over). We also examined the extent to which purely local (i.e., trial-to-trial) interactions, rather than long-range (i.e., over the course of full blocks of trials), could explain the RT differences. Given that the PCE is inherently a manifestation of experience, and that the presence/absence of feedback is likely to modulate the impact of long-term experience (e.g., learning), in Experiment 1 we investigated the issues above in the context of the methodological approach (no feedback) that is perhaps more common in broader attention and cognition research. In Experiment 2 we then included feedback into our design to more closely align with typical studies of the PCE.

Overarching Methodological Approach Across Experiment 1 & 2

Overview of procedures and sample

Participants (Experiment 1 n = 54; Experiment 2 n = 60; Gender: 60.8% Female; Race: 21.6% Asian; 62.2% White; 16.2% Other or Multiple) were recruited from an introduction to psychology participant pool and given course credit for their participation. All participants provided informed consent, and all procedures were approved by the University of Wisconsin-Madison Institutional Review Board.

Details of task

Participants completed 2 blocks of an arrow flanker task (B. A. Eriksen & Eriksen, 1974; Fan et al., 2002; Rueda et al., 2005). Each block consisted of 400 trials that could be compatible (target facing the same direction as flankers) or incompatible (target facing the opposite direction as flankers; see Figure S1). Each block included one of three possible predetermined ratios of compatible to incompatible trials: 20-80, 50-50, or 80-20. Participants were pseudorandomly assigned to conditions such that each participant completed blocks containing two different compatibility proportions (i.e., task statistics).

The flanker task was run in Python using the Psychopy library on a 22 inch Dell monitor in a dimly lit room. Participants sat approximately 60 cm from the monitor. Stimuli consisted of arrows overlaid on cartoon fish (Cochrane et al., 2019; Rueda et al., 2005) 1.5 degrees of visual angle wide and placed with centers 1.65 degrees apart (i.e., .15 degrees separating stimuli). The center fish was always presented at the center of the screen. Each block included stimuli of a single randomly chosen color of light purple, green, or orange (see Supplemental Information for stimuli). A 100 millisecond centrally-located cross cue was first presented, after which there was a blank screen for a random time between 100 and 300 milliseconds prior to stimulus onset. Responses were recorded on a standard keyboard by pressing the arrow key corresponding to the target stimulus (i.e., left or right). After response, an 850 ms delay occurred.

Analyses

Incorrect trials were first excluded (5.22% across all participants). Because error rates were very low, no analysis of errors was conducted. Trials with RT over 1.5 seconds (0.46%) or below .2 seconds (0.58%) were then excluded. Where appropriate we use linear mixed-effects models, with participant-level random intercepts, using the R package **lme4** (Bates et al., 2015, p. 4) with degrees-of-freedom approximation using the package **pbkrtest** (Halekoh & Højsgaard, 2014). Proportion compatible was treated as a three-level categorical variable, allowing for asymmetric effects of proportions above or below 50%. In tests of the overall PCE, the reference proportion was set to be the 20% condition in order to simplify our confirmation of the monotonically increasing flanker effect with increasing proportion of compatible flankers. In all other tests the reference proportion was set to be the 50% condition, providing for a clear interpretation of possible asymmetric effects. We reported overall PCE results as interactions between current-trial block proportion compatible. When testing changes in the flanker effect we first averaged the RT for each participant's compatible and incompatible trials separately for the first 50 trials and last 50 trials of each block. This provided us with 8 mean RTs per participant for subsequent analyses. Change in mean RT was next calculated for each participant's blocks by subtracting compatible

trials from incompatible trials (i.e., flanker effect) and subtracting the final-50-trial flanker effect from the first-50-trial flanker effect. Results report differences in these change scores.

Experiment 1 Results

Was the canonical PCE observed?

The PCE is typically observed as an increase in the magnitude of the flanker effect in conditions with more compatible trials (or likewise, a decrease in flanker effect in conditions with more incompatible trials). We first tested for this pattern across all data to provide estimates averaging over proportions compatible and block number. We fit a linear mixed-effects model predicting RT with main effects and the interaction of trial incompatibility and proportion compatible in a block, using 20% compatible as a reference, while controlling for the random effect of participant-level mean RT (see Table 1). The predicted PCE would manifest as negative coefficients for higher proportions compatible (i.e., faster RT on compatible trials) along with positive coefficients for the interaction between trial type and higher proportions compatible (i.e., slower RT on incompatible trials).

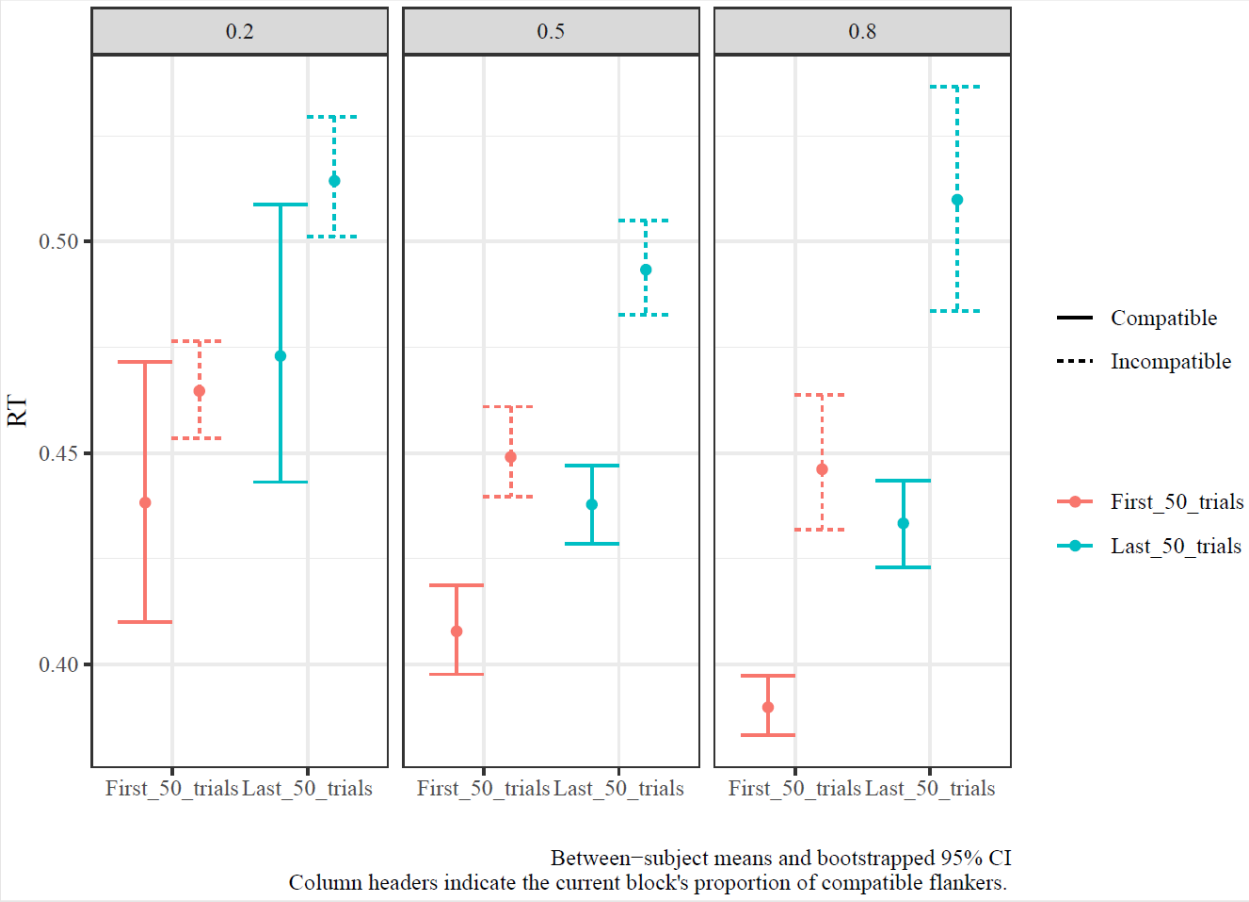
Table 1. Entire-block proportion flanker effect on compatible and incompatible trials

	Estimate	Std. Error	t value
(Intercept)	0.439	0.010	44.095
50%_compat	-0.012	0.003	-4.161
80%_compat	-0.012	0.003	-4.070
IsIncompatible	0.040	0.003	15.269
50%_compat:IsIncompatible	0.008	0.004	2.344
80%_compat:IsIncompatible	0.018	0.004	4.828

Consistent with the PCE, on average, relative to a block with 20% compatible trials, blocks with more compatible trials had reliably lower compatible-trial RTs and higher incompatible-trial RTs (and thus a larger flanker effect; see Figure 1). In examining the fit model coefficients, the effect on compatible trials appeared to have saturated by 50% compatible while the effect on incompatible trials continued to increase through 80% compatible.

Figure 1. Block 1 RT in the absence of feedback, separated by the first 50 and last 50 trials. Column headers indicate current block's proportion compatible. Several outcomes are evident in this plot. First, the expected PCE is seen (i.e., the difference between RTs for compatible and incompatible trials is

smaller in the 20% compatible condition as compared to the 80% compatible condition). Second, RTs through time generally increase (i.e., the difference between RTs during the first 50 trials and the last 50 trials). Between-subjects means and 95% CI are indicated. See Figure S2 for full pattern of results across both blocks and experiments.



Does the flanker effect reliably change over the course of participants' first 400 trials?

As noted in the Introduction, when stimulus/response pairs are arbitrary (e.g., if the two possible targets are a square and a diamond, indicated by pressing the "Z" or the "M" buttons on the keyboard), the flanker effect would seemingly require some degree of learning (i.e., to learn that some stimulus/response pairs are "incompatible" requires an understanding of the task). Yet, in the case of an arrow-based task with natural key mappings, it is less clear whether task-based experience is necessary to observe the flanker effect. Indeed, university students presumably enter such a task with a great deal of experience with the idea of left/right arrows as incongruent/opposite one another. Furthermore, as noted in the introduction, the methodological approach used in Experiment 1 (utilizing no feedback) is often employed with the goal of eliminating, or at least reducing, change through time (e.g., learning has the

potential to be particularly problematic in research that makes use of the same task across multiple time points).

To examine whether we could detect systematic change over the course of participants' first 400 trials, we tested change in incompatible and compatible RTs using a linear mixed-effects model predicting change in RT with a fixed effect of trial compatibility and a random intercept for each participant. In this model, participants' flanker effect within their first block did change reliably, although not due to a canonical learning effect (e.g., where RTs become faster through time and flanker effects decrease). Instead, the flanker effect reliably increased rather than decreased (b = 0.016, T = 2.283, p = 0.027). When analyzed separately, reliable change was evident in both incompatible trials (b = 0.05, T = 4.02, $p < .001$) and compatible trials (b = 0.034, T = 3.16, p = 0.003) with the increase in flanker effect noted above being driven by a further significant increase in incompatible RT relative to compatible RT. Therefore, the results were inconsistent with the proposition that flanker effects are stable through a block when highly familiar stimuli (arrows) are utilized without feedback. While we observed reliable changes in the flanker effect in participants' first 400 trials, these changes were also not in the direction that would be expected from learning, as rather than becoming faster through time, participants' RTs slowed through time.

Does compatibility proportion influence within-block changes in response time?

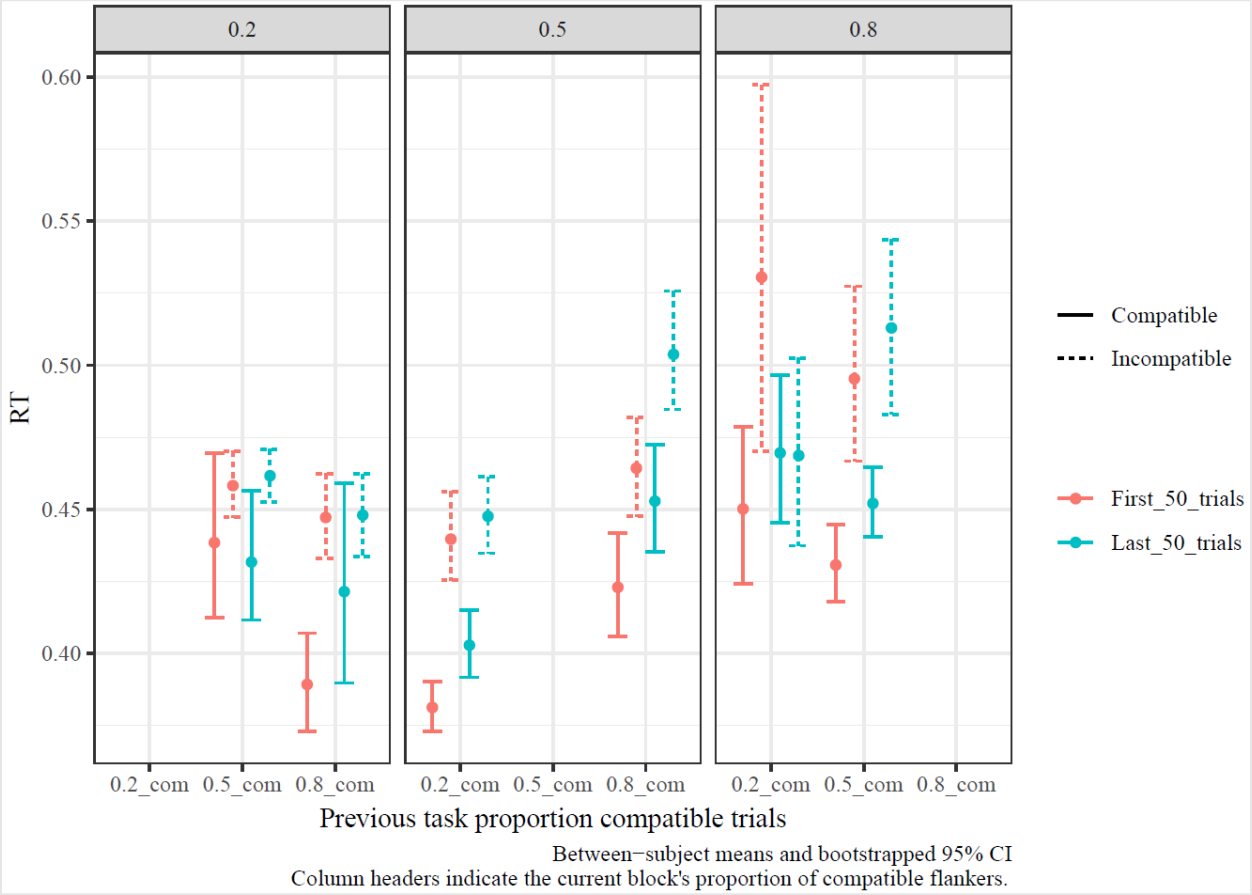
In our first two analyses we observed canonical PCE effects in participants' initial block, with an associated increase in RT and flanker effect across that block. One reason for these patterns may have been that the proportion of compatible trials differentially affected the change in RT that occurs through time (e.g., differential increases in RT leading to selectively larger flanker effects in certain conditions). As such, we next tested whether differences in RT changes through a block, as a function of the proportion compatibility in the block, explained the overall PCE. We fit a linear mixed-effects model predicting each participant's change in RT from the first 50 to the last 50 trials with the main effects of trial compatibility and block proportion compatible as well as their interaction. The model also included participant-level random effect intercepts. There were no reliable influences of a high or low proportion of compatible trials, compared to 50% compatible, on the change in flanker effect (both $|T| < 0.62$), incompatible trials (both $|T| < 0.69$), or compatible trials (both $|T| < 0.6$). Despite the clear overall differences between blocks with different proportion of compatible trials (see Figure 1), this effect was not explained by comparisons between the flanker effects in the first 50 trials and last 50 trials. Instead, measured in this way, the change in flanker effect through time between conditions was indistinguishable. Given that the PCE must arise in response to experience with the task, the present analysis suggests that

the changes in RT associated with the PCE likely occur quickly (e.g., within the first 50 trials; see *Finer time-scales* section below; cf. Abrahamse et al., 2013).

Does previous experience with one block of the task affect subsequent within-block change in flanker effect?

In previous work, Abrahamse and colleagues (2013) observed asymmetric carry-over effects in the PCE using a Stroop task. As such, in our next set of analyses we tested whether, when compared to first blocks with the same compatibility proportion, RT changes in subsequent blocks of trials are influenced by prior task experience. Figure 2 shows the distributions of these second-block RTs separated by participants' first blocks. We predicted that prior experience with a high-compatibility block would mean that participants would begin their second block with a relatively large flanker effect, thereby leading to a disproportionate decrease in flanker effect across the second block. The opposite pattern was predicted for blocks following the low-compatibility first blocks.

Figure 2. Block 2 RT in the absence of feedback, separated by the first 50 and last 50 trials. As with participants' first blocks, RT increases in several cases. When the first block had 20% compatible trials, second block was more likely to show a pattern of converging compatible and incompatible RT. Column headers indicate current block's proportion compatible; x-axes indicate prior block's proportion compatible. Between-subjects means and 95% CI are indicated.



We fit a linear mixed-effects model to change in flanker effect (i.e., mean of flanker effect on last 50 trials minus mean of flanker effect on first 50 trials) with fixed effects of previous block compatibility and current block compatibility while including by-subject random intercept (see Table 2). The significant *Intercept* term in this model indicates an overall increase in flanker effect over the course of a first block of flanker trials in blocks with equal numbers of compatible and incompatible trials. The same pattern is very similar in first blocks with 20% compatible but is non-significantly attenuated in first blocks with 80% compatible. Further, the reliable increase in flanker effect evident in the *Intercept* effect is significantly attenuated and fully reversed, when a second block is preceded by a first block with many incompatible trials.

Table 2. Linear mixed-effects model predicting change in flanker effect for feedback-absent participants, accounting for subject-level intercepts. Intercept indicates change in flanker observed on first block on blocks with equal numbers of compatible and incompatible trials.

	Estimate	Std. Error	t value	p_value
(Intercept)	0.022	0.011	2.104	0.038

previous_20%	-0.046	0.017	-2.778	0.007
previous_50%	-0.008	0.015	-0.523	0.603
previous_80%	-0.021	0.017	-1.203	0.234
20%_Compat	-0.003	0.015	-0.181	0.858
80%_Compat	-0.021	0.015	-1.415	0.163

In order to clarify these effects we separately examined within-block changes in incompatible trials and compatible trials while controlling for the current block's compatibility proportion (see Tables 3 and 4). The comparison of interest for these changes in RT for incompatible or compatible trials was whether the change is different after certain types of blocks, as opposed to participants' first block. That is, the *Intercept* parameter indicates first-block change in RT (controlling for compatibility proportion).

Table 3. Linear mixed-effects model predicting change in incompatible-trial RT, accounting for subject-level intercepts. Fewer compatible trials in the first block systematically led to a smaller increase in incompatible-trial RT in the second block.

	Estimate	Std. Error	t value	p_value
(Intercept)	0.054	0.016	3.467	0.001
previous_20%	-0.068	0.024	-2.786	0.007
previous_50%	-0.047	0.022	-2.109	0.038
previous_80%	-0.024	0.025	-0.951	0.346
20%_Compat	-0.007	0.021	-0.323	0.749
80%_Compat	-0.006	0.021	-0.290	0.774

While an increase in incompatible-trial RT was reliable for the first block (Intercept), the associated change of incompatible RT in the second block was attenuated only when the first block included relatively few (50% or 20%) compatible trials.

Table 4. Linear mixed-effects model predicting change in compatible-trial RT, accounting for subject-level intercepts.

	Estimate	Std. Error	t value	p_value
(Intercept)	0.032	0.013	2.377	0.020
previous_20%	-0.022	0.021	-1.042	0.303
previous_50%	-0.039	0.019	-2.050	0.044
previous_80%	-0.003	0.021	-0.153	0.879

20%_Compat	-0.004	0.018	-0.233	0.817
80%_Compat	0.014	0.018	0.790	0.435

While an increase in compatible-trial RT was reliable for the first block (Intercept), the associated change of compatible RT in the second block was significantly attenuated only when the first block included an intermediate number (50%) of compatible trials. This coefficient for this intermediate block was numerically more similar to the low-compatibility first block than the high-compatibility first block.

Experiment 1 Discussion

Experiment 1 tested for the presence of learning giving rise to the PCE in initial and subsequent blocks of an arrow flanker task in the absence of feedback. First, consistent with previous work, in Experiment 1 we found evidence for a PCE. The magnitude of the flanker effect was smaller in blocks with larger numbers of incompatible trials. Second, while we observed that the final magnitude of the flanker effect did emerge through time, the direction of this change through time was not consistent with a learning effect. Instead, within the first block there was a tendency toward ever slower RTs – with this effect being magnified in incompatible trials. Furthermore, we found that this drift toward slower RTs was similar across percent compatible conditions, and thus differences in a slow drift in RTs across conditions did not appear to explain the PCE. Finally, we found reliable differences in RT change in participants' second blocks as a function of their experiences in their first blocks. Specifically, while there was an overall pattern of incompatible RTs increasing over the course of participants' blocks, more previous experience with incompatible trials led to a subsequent attenuation of within-block RT change in the second block. This, in turn, led to the overall within-block increases in flanker effect being significantly reversed in blocks following a 20%-compatible block. More previous experience with incompatible trials also led to an attenuation of the within-block increase in compatible-trial RT, but this effect was strongest for participants who completed 50%-compatible blocks first.

Experiment 2 Results

While Experiment 1 was a clear demonstration of cross-block patterns of RT change, the overall pattern of increasing RTs was inconsistent with the changes in behavior that usually come with experience (Newell & Rosenbloom, 1981; Wenke et al., 2015). That is, while participants clearly were adjusting their behavior in the task through time, these adjustments did not manifest as a decrease in overall RT (i.e., as might be expected from learning). One possible reason for the failure to observe a decrease in RTs through time may be a lack of feedback helping participants to modulate their behaviors. In Experiment 2 we thus implemented the same methods as in Experiment 1 with the exception that we

provided participants with both response time and accuracy feedback. After each trial, participants saw a screen with their response time on the previous trial. The text was typically blue, but was colored red if the response was incorrect or longer than 500 ms. This feedback was presented for the first 750 ms of the total 850 ms inter-trial interval (see Methods).

Was the canonical PCE observed?

As with Experiment 1, we first tested for the overall expected positive interaction between trial compatibility and proportion compatible, across blocks and proportions congruent and accounting for participant-level random-effect intercepts. This predicted PCE should manifest as negative coefficients associated with higher proportions compatible (i.e., faster compatible RT) along with positive coefficients for the interaction between trial type (i.e., incompatible RT) and higher proportions compatible. When compared to a block with 20% compatible trials, increasing the proportion of compatible trials reliably decreased RT on compatible trials while increasing RT on incompatible trials (see Table 5). This led to the PCE, an increase in the flanker effect in conditions with a larger proportion of compatible trials. Thus, at this broad level of analysis we replicated the results of Experiment 1. Unlike Experiment 1 though, the magnitude of both effects continued to increase from 50% compatible to 80% compatible.

Table 5. Entire-block proportion flanker effect on compatible and incompatible trials.

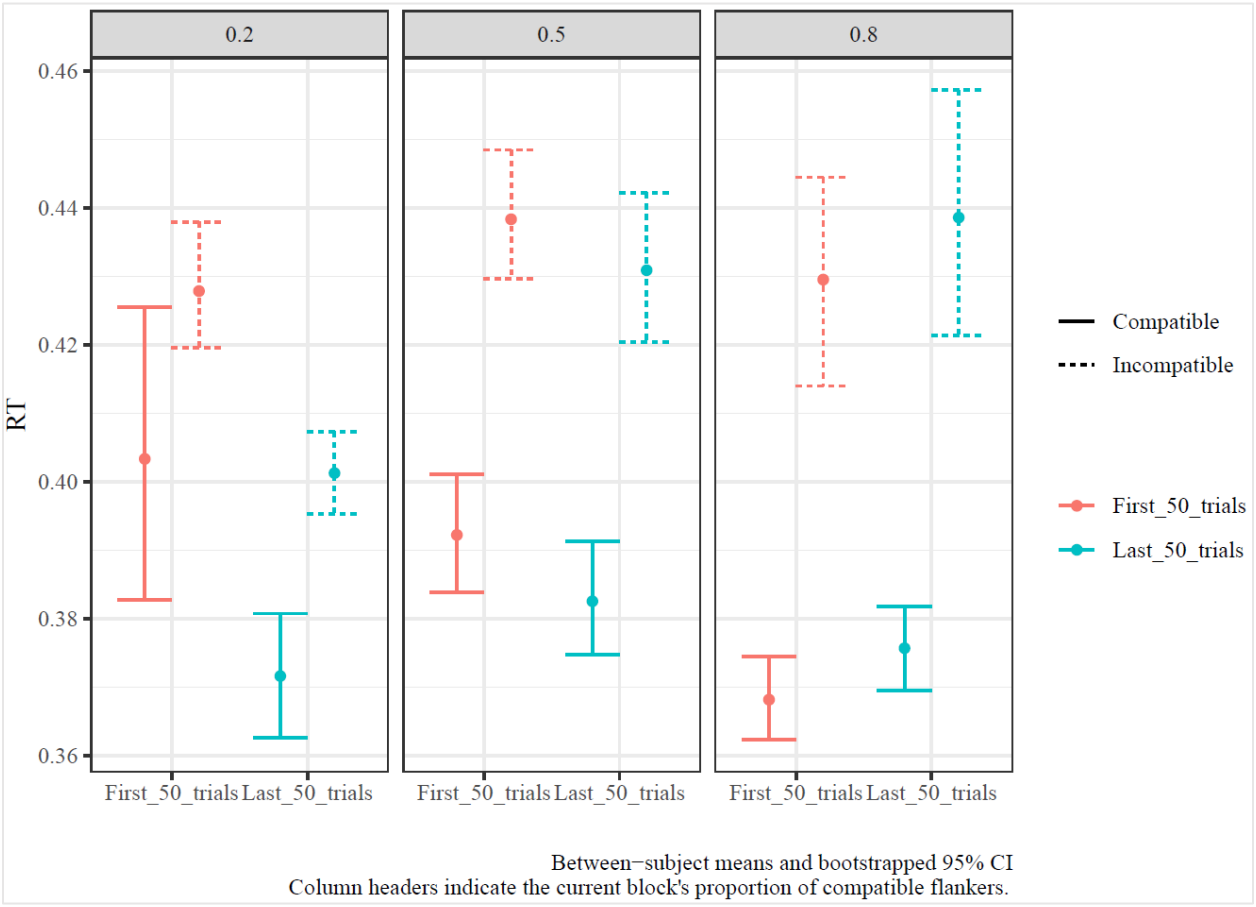
	Estimate	Std. Error	t value
(Intercept)	0.380	0.005	76.484
50%_compat	-0.009	0.002	-4.692
80%_compat	-0.015	0.002	-8.450
IsIncompatible	0.035	0.002	20.917
50%_compat:IsIncompatible	0.005	0.002	2.450
80%_compat:IsIncompatible	0.020	0.002	8.008

Does the flanker effect reliably change over the course of participants' first 400 trials?

In Experiment 1, we observed a reliable change in both RT and the flanker effect within the first 400 trials. However, the direction of this change was such that participants got progressively slower throughout the block. We surmised that this may have been due to the lack of explicit trial-by-trial feedback. We thus predicted that the introduction of feedback would, at a minimum, eliminate the drift toward slower RTs throughout the first block. We additionally expected a decrease in RT for both compatible and incompatible trials in the first block, independently of proportion compatible (as would be expected by learning). Further, disproportionate learning on incompatible trials should lead to a reduction

in flanker effects from the beginning to the end of the block. We tested for these overall effects using a linear mixed-effects model predicting change in RT in participants' first blocks with the fixed effect of trial type while controlling for participant-level intercepts.

Figure 3. Block 1 RT in the presence of feedback, separated by the first 50 and last 50 trials. Unlike in Experiment 1, there is no overall trend toward slower RT on later trials. Because changes for compatible trials were similar to changes for incompatible trials within blocks of the same proportion compatible, there was no reliable change in flanker effect. Column headers indicate current block's proportion compatible. Between-subjects means and 95% CI are indicated.



Participants exhibited no change in flanker effect in their first block (b = -0.001, T = -0.157, p = 0.875). There was likewise no overall reliable change in either incompatible trials (b = -0.009, T = -0.44, p = 0.658) or compatible trials (b = -0.009, T = -0.51, p = 0.614). Thus, while the increases in RT we observed in Experiment 1 were no longer evident in the presence of feedback, participants in Experiment 2 did not demonstrate learning (i.e., reduction of RT).

Do block compatibility proportions influence changes in response time?

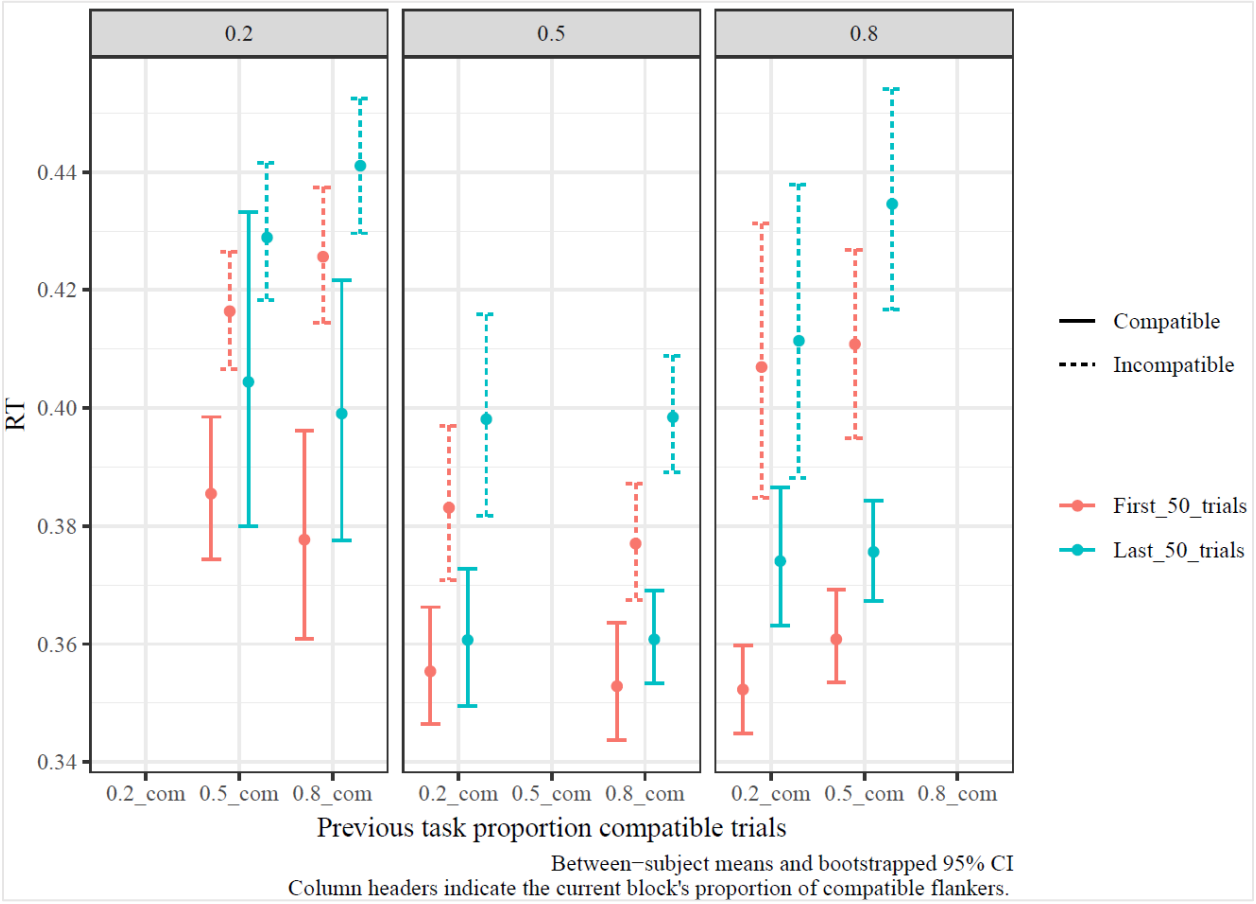
One possible reason for a lack of overall change in RT in the first block may be that certain combinations of compatibility proportions are more conducive to learning than others. Specifically, the overall PCE led us to predict that smaller proportions of compatible trials may be more associated with decreases in RT and flanker effect. This would be evident in conditions with fewer compatible trials having reliably larger decreases (i.e., more negative coefficients) in RT and flanker effect from the beginning of blocks to the end. In participants' first blocks, there was no reliable influence of high or low proportions compatible trials, compared to 50% compatible, on the change in flanker effect (both $|T| < 0.39$), incompatible trials (both $|T| < 0.98$), or compatible trials (both $|T| < 0.89$). While the numerical patterns of RT increases and decreases appeared consistent with our predictions (see Figure 3), these between-group differences in RT change did not lead to statistically reliable differences when 80%-compatible or 20%-compatible blocks were compared to 50%-compatible blocks. As with Experiment 1 we found reliable between-condition PCE when testing all trials, but our analyses of the first and last 50 trials' RT were insensitive to changes in flanker effect giving rise to the PCE (see *Finer time-scales* section for further tests of these changes).

Does previous experience with a task alter within-block change in flanker effect?

Despite the lack of evident learning when comparing the first 50 and last 50 trials of participants' first blocks, it remains possible that task learning may have carried over and interacted with performance on the second block. Specifically, if first-block learning-induced initial compatibility-effect changes in subsequent blocks, this may have provided exaggerated or depressed development of the PCE in the second blocks (see Figure 4).

Figure 4. Block 2 RT in the presence of feedback, separated by the first 50 and last 50 trials. There was no reliable trend of learning (i.e., decreasing RT) observed. Column headers indicate current block's proportion compatible; x-axes indicate prior block's proportion compatible. Between-subjects means and

95% CI are indicated.



We tested for the effects of prior compatibility experience by predicting the change in flanker effect (i.e., mean on last 50 trials minus mean on first 50 trials) with blocks' proportion compatible and the previous proportion compatible (i.e., none/*Intercept* if predicted data was from the first block, or the first block's proportion of the predicted data was from the second block).

Table 6. Linear mixed-effects model predicting change in flanker effect for feedback-present participants, accounting for subject-level intercepts. Intercept indicates change in flanker observed on first block.

	Estimate	Std. Error	t value	p_value
(Intercept)	0.005	0.009	0.578	0.566
previous_20%	-0.001	0.012	-0.114	0.910
previous_50%	0.005	0.012	0.435	0.666
previous_80%	0.003	0.012	0.208	0.836
20%_Compat	-0.014	0.011	-1.277	0.208
80%_Compat	-0.003	0.011	-0.275	0.786

Changes in flanker effect were not reliable. We did not find any systematic differences in changes in flanker effect due to prior experience or due to current proportion compatible (each controlling for the other). Next, we separately tested for within-block changes in incompatible trials and compatible trials while controlling for the current block's compatibility proportion. Note the negative *Intercept* coefficients, indicating a trend toward overall improvements during the first block (see Figure 3).

Table 7. Linear mixed-effects model using prior task compatibility proportions to predict change in incompatible-trial RT, accounting for subject-level intercepts.

	Estimate	Std. Error	t value	p_value
(Intercept)	-0.009	0.014	-0.608	0.547
previous_20%	0.017	0.019	0.913	0.367
previous_50%	0.032	0.020	1.620	0.111
previous_80%	0.033	0.020	1.682	0.098
20%_Compat	-0.022	0.018	-1.256	0.216
80%_Compat	0.014	0.018	0.808	0.426

Table 8. Linear mixed-effects model using prior task compatibility proportion to predict change in compatible-trial RT, accounting for subject-level intercepts.

	Estimate	Std. Error	t value	p_value
(Intercept)	-0.015	0.013	-1.132	0.262
previous_20%	0.020	0.018	1.126	0.265
previous_50%	0.025	0.018	1.386	0.171
previous_80%	0.031	0.018	1.734	0.088
20%_Compat	-0.006	0.016	-0.342	0.735
80%_Compat	0.017	0.016	1.042	0.304

Neither compatible nor incompatible RT demonstrated change that was reliably affected by current or previous proportion compatible trials. The patterns of RT increase or decrease, although not statistically reliable, can also be observed in Figure 4.

Experiment 2 Discussion

Experiment 2 tested for learning-related changes in the PCE in two blocks of an arrow flanker task feedback when feedback was provided. First, as with Experiment 1, in Experiment 2 we replicated

the canonical PCE effect when implementing the flanker effect with feedback. Yet, while the magnitude of flanker effect was different when comparing blocks with different task statistics, our measurements of change were insensitive to any within-block changes in RT (i.e., comparing the means of the first vs. last 50 trials). This stands in contrast to Experiment 1 wherein we found systematic increases in RT that accompanied the overall PCE. Experiment 2 instead provided evidence for relatively stable response time distributions for both compatible and incompatible trials when participants were provided with feedback. The fact that the PCE was observed both in an experiment (Experiment 1) where overall RTs were drifting upward across a block and also where overall RTs were reasonably stable across a block (Experiment 2), means that it remains unclear how condition-level differences in the PCE arise through time. In the next section, we explore several analyses that are more sensitive to changes in RT on timescales that the previous analyses may not have been able to capture.

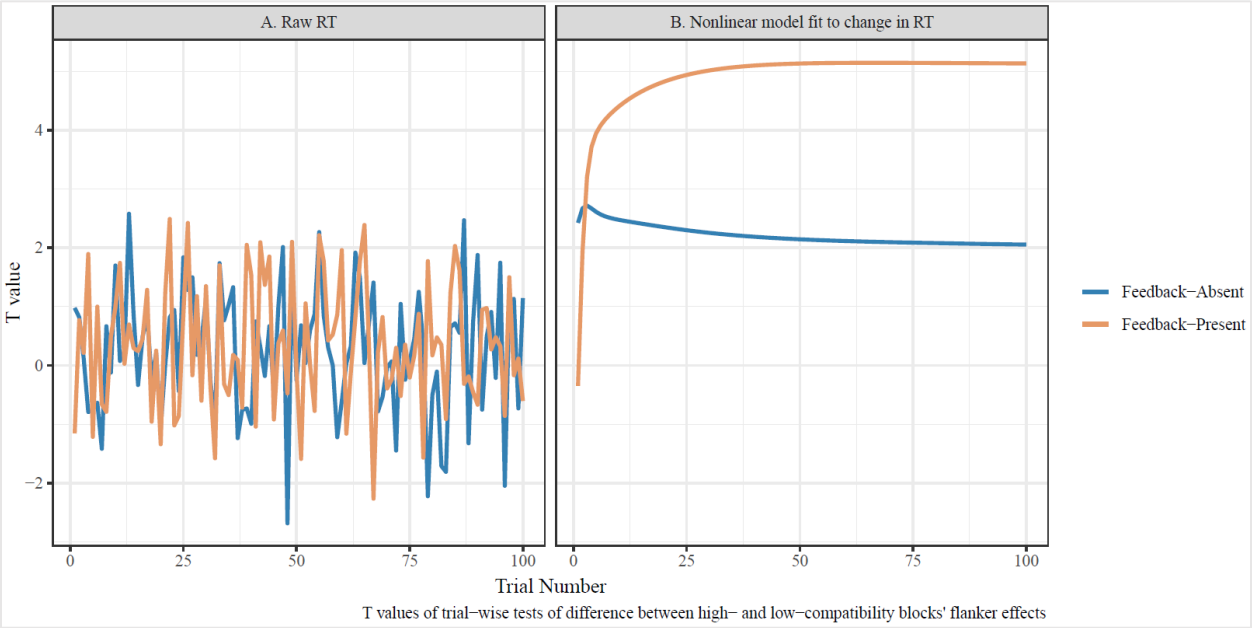
Evaluating how the PCE emerges over the course of experimental blocks on finer time-scales

In two experiments, we found robust evidence for the influence of compatibility proportion on overall response times, with less evidence for systematic within-block changes in response times. This is counter-intuitive because such an effect must necessarily be a function of time. In other words, participants do not have prior knowledge of the upcoming task statistics, so their response on the very first trial of a block cannot be impacted by the distribution of trials to come. Thus, at a minimum, the effects must evolve after the very first trial. Yet, despite prior research indicating the continuation of this development over the course of hundreds of trials (Abrahamse et al, 2013), our comparisons of blocks' first 50 trials to the last 50 trials were insensitive to the development of the PCE. Furthermore, the presence of the PCE both in Experiment 1, where within blocks we observed overall slow increases in RT, and Experiment 2, where no such changes were observed, suggests the need to look at other time-scales. Given this, we conducted a set of analyses using more time-sensitive measures of the development of the PCE. We first iteratively tested the trial number on which the flanker effect differs by proportion compatible when predicting RTs. Due to the noise inherent in this analysis, we next used a parametric model of RT change (i.e., assuming monotonic change) in similar analyses. Last we tested whether our results were accounted for by contingencies on a trial-to-trial timescale.

To examine RTs at a finer time-scale, we first iteratively fit mixed-effects models predicting raw RTs on the first trial, second trial, etc. These models included fixed effects of current block proportion compatible, current trial proportion compatible, and the interaction between these two. The models also controlled for previous block proportion compatible and participant-level intercepts. Low-compatibility

blocks had flanker effects that diverged from the high-compatibility block quickly, with the flanker effect appearing different by block-type (i.e., interaction between block compatibility and trial compatibility; T > 2) by trial 13 in Experiment 1 and trial 22 in Experiment 2 (see Figure 5). This is consistent with the idea that these effects appear quite quickly in a block. Yet, we note that these estimates also appear to be heavily influenced by sampling noise. The reported test would likely have a false positive rate of .05, leading to approximately 1 out of 20 tests appearing reliable by chance.

Figure 5. Time course of change in flanker effect between low-compatibility and high-compatibility conditions over the first 100 trials, as tested by linear mixed-effects models fit to each of the first 100 trial numbers. (A) When testing differences in raw RT, differences are noisy between high- and low-compatibility blocks. (B) When testing differences in RT fit by nonlinear regression, the rapid differences between conditions are clearer.



To mitigate noise in this analysis, we conducted the same iterative tests on nonlinear model predictions of the flanker effect. We first fit an exponential learning model to each participant's flanker effect for each block using the R package **TEfits** (Cochrane, 2020), producing by-trial estimates of the flanker effect (see Supplementary Information for model details). We used these model estimates in an analysis using the same iterative by-trial models described above, and again found differential flanker effects emerging very quickly in both experiments (in this analysis, before trial #4, see Figure 5). Thus, in analysis of raw data as well as data fit with learning models, these timescales are clearly shorter than that necessary to effectively learn from task compatibility proportion (i.e., there is little possibility of learning

that a block has 80% compatible trials in only 4 total trials). Further, when testing for differences between the parameters themselves of the nonlinear models (i.e., starting flanker effect, rate of change, or asymptotic flanker effect), no reliable effects were evident of current or previous proportions compatible. This indicates that, like aggregating over the first and last 50 trials, by-trial learning models reported here are unable to fully capture the timescale of change associated with the PCE.

As noted in the introduction, another possible root of the block-wise PCE is not necessarily based upon learning *per se*, but simple trial-to-trial contingencies in behavior. For example, flanker effects may be smaller after incompatible trials (Braem et al., 2019; Davelaar & Stevens, 2009; Gratton et al., 1992, cf. Duthoo et al., 2014). The PCE would then emerge because there are simply more incompatible trials in some blocks. Using linear mixed-effects models to test the interaction between current and previous-trial compatibility, we found that this effect does exist overall in both feedback-absent ($b = -0.011$, $T = -3.96$) and feedback-present conditions ($b = -0.009$, $T = -5.21$).

Interestingly though, the links between these rapid trial-wise changes in behavior and longer within-block or cross-block effects remained unclear. For instance, in mixed-effects models for each experiment's first block predicting RT with previous trial's compatibility, current trial's compatibility, overall proportion compatible, and second-order interactions between current trial compatibility and the other predictors, the interactions between block proportion compatible and current trial compatibility generally remained reliable even when controlling for the other effects (with 50% compatible as a reference; without-feedback 20% $T = -1.25$, 80% $T = 2.56$; with-feedback 20% $T = -3.2$, 80% $T = 3.1$). These results show that decreased (20% - compatibility blocks) or increased (80% - compatible blocks) flanker effects, when compared to the 50%-compatible block, are not fully accounted for by trial-to-trial contingencies. Thus, while RTs are reliably influenced by the previous trial's compatibility, the PCE was still evident in our data when controlling for this short-term effect.

General Discussion

Here we tested the time course of adaptation of attentional control to task statistics. Following Eriksen (C. W. Eriksen, 1995), our findings broadly support a view of visual selective attention as rapidly contextually modulated via scaling inhibition. First, we replicated the canonical *Proportional Compatibility Effect* (PCE) across two experiments using an arrow flanker task, the first without feedback and the second with feedback. In each experiment, RTs on incompatible trials were slower in blocks with

fewer incompatible trials. There was a corresponding effect on compatible trials, wherein RTs were longer in blocks with fewer compatible trials.

However, some patterns diverged when examining the temporal dynamics of performance in these two experiments. In Experiment 1, which did not include explicit performance feedback, participants' RTs reliably increased over the course of the first block of trials. This increase runs contrary both to the prediction of a low-level learning model (i.e., where RTs should generally decrease with experience with a task), as well as one of the primary justifications for omitting feedback in the broader attention and cognition literature (i.e., that omitting feedback would produce more stable behavior). Instead participants' first feedback-absent blocks were associated with significant increases in incompatible-trial RT and in compatible-trial RT, leading to a significant increase in flanker effect over time. In contrast, when feedback was present, no changes in RT or flanker effect were reliably evident when comparing the first 50 trials to the last 50.

Further, in Experiment 1, patterns of RT change in sequential blocks showed that experience with more incompatible trials in an initial block led to a reliable attenuation of flanker-effect increase in subsequent blocks. That is, when participants completed an initial block with only 20% compatible trials, their second block's flanker effect was likely to decrease over time rather than increase. This effect was due largely to RTs in incompatible trials; within-block increases in incompatible-trial RT were significantly attenuated when blocks were preceded by 20% or 50% compatible blocks. The magnitude of this attenuation was approximately linear across 20%, 50%, and 80% compatible blocks. Compatible-trial attenuation of RT increase follows a different and nonmonotonic pattern in which only the effect of being preceded by a 50% compatible block is significant.

In all, the results of Experiment 1 indicate global (i.e., above and beyond local statistics) learning from experience with incompatible trials such that second blocks' incompatible trials are improved with, but not without, majority-incompatible initial blocks. This learning cannot be wholly explained by carry-over from block 1 to block 2 – if second blocks simply demonstrated initially lowered RT for incompatible trials due to previous experience with many incompatible trials, RT would not be expected to decrease even further within these second blocks. In other words, second-block decreases in incompatible RT seem to attenuate the PCE rather than being a result of the PCE. However, each of these interpretations is clouded by the overall increases in RT observed in Experiment 1. In RT measures learning is typically considered to be manifested in the form of decreased RT. There is no necessity in this relationship, though, and it is clearly possible that any task may be associated with learning which does

not influence the measured behavior (i.e., measurements may be inadequate indicators of internal states). By testing sequential learning, we were able to observe learning effects that were not evident in within-block measures.

Despite replicating the canonical PCE effect, the patterns of learning in Experiment 1 were unexpected given standard theories of learning (i.e., locally increasing RT with learning-related attenuation rather than overall decreasing RT). As such, it may be possible that our pattern of results in Experiment 1 arose from processes of change that are not typical of PCE results. One reason for the differences may have been that Experiment 1, unlike most PCE research, involved the use of flanker tasks without feedback. PCE is most often studied in Simon-like or Stroop-like tasks (e.g., Hutchison, 2011; Spinelli et al., 2019; Wühr et al., 2015) and/or using feedback (e.g., Lehle & Hübner, 2008; Wendt et al., 2008). It is possible that the error signals regarding incorrect or slow trials, whether explicit or self-monitoring, that participants receive in a no-feedback flanker task are weaker than the more commonly studied paradigms and are therefore unable to drive learning in the form of overall decreases in RT. In this context *error* would be any noncompliance with the instructions to complete the task quickly and accurately. The mechanism by which the error signal would act (e.g., narrowed scope of attention, increased engagement) is not specifically of interest here. Instead, we simply wanted to align Experiment 2 with a set of theories that assumes that participants receive strong feedback signals. To investigate the possibility that increased error signal would lead to more canonical patterns of RT change over time, we tested the sensitivity of the observed learning to the inclusion of by-trial feedback.

In Experiment 2 we found that, unlike in Experiment 1, RT on compatible trials as well as RT on incompatible trials decreased over the course of each block. However, neither of these effects was reliable. Compatible RT decreased more than incompatible RT, leading to a non-reliable increase in flanker effect. Apart from the overall PCE, no statistically reliable effects were observed in any of the Experiment 2 data. This lack of reliable effects is remarkable given the robustness of the PCE; proportion-flanker effects must necessarily develop over time with accumulated experience with task statistics, yet the measures used here are insensitive to these changes. We next used by-trial nonlinear regressions of the flanker effect fit to each block of each participant. In comparing the coefficients and predictions from these models, we still found no reliable condition differences in changes over the course of blocks. This indicates that the evident change in RT distributions must have occurred rapidly enough that even by-trial learning models are unable to effectively capture the change. The PCE remains reliable even when controlling for adjacent-trial effects, however, indicating that the PCE we observed has a source above and beyond single-trial fluctuations.

Here we have taken an approach to the study of attention and that is quite different than most PCE studies. In two experiments we replicated the PCE effect when averaged across all trials as well as within-trial pairs (i.e., smaller flanker effects on trials following incompatible trials). However, given the overall goal of integrating PCE effect into a learning framework that would be applicable to broader fields of attention and cognition, our results are mixed. As with PCE results more generally, the systematic changes in flanker effect we observed should act as a contextualizing caution to researchers putting a large stake on single measures of response competition. Our reported divergence between feedback-present and feedback-absent experiments should provide a basis for future research to consider the combination of feedback and compatibility best suited to answer questions of response competition, individual differences, learning, or fatigue. In addition, the possible influence of sequential task demands cannot be dismissed even in this fairly simple task.

Conclusion

Our work corroborates the broader PCE literature in providing unequivocal evidence for interacting bottom-up (i.e., stimulus-driven) and top-down (e.g., learning, attentional modulation) processes in response competition. Despite our goals of identifying sources of variation in the magnitude and timescale of PCE modulations, our measures were insensitive to the time course of learning. In contrast with previous work positing the development of the PCE over hundreds of trials (Abrahamse et al., 2013), we found the presence of the PCE very early when using by-trial modeling of response times. This indirectly supports a timescale of modulation as small as single trials. Nonetheless, future work would benefit from methodological innovations facilitating an identification of the interacting timescales giving rise to sequential PCE effects.

Funding

This work was supported in part by Office of Naval Research Grant ONR-N000141712049. This funding source had no direct involvement in study design, data collection, analysis, manuscript preparation, or any other direct involvement in this research.

Open Practices Statement: *The data and materials for all experiments are available at 10.17605/OSF.IO/KMBZ8 and none of the experiments were preregistered*

References

Abrahamse, E., Braem, S., Notebaert, W., & Verguts, T. (2016). Grounding cognitive control in associative learning. *Psychological Bulletin*, *142*(7), 693–728. https://doi.org/10.1037/bul0000047

Abrahamse, E., Duthoo, W., Notebaert, W., & Risko, E. F. (2013). Attention modulation by proportion congruency: The asymmetrical list shifting effect. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *39*(5), 1552–1562. https://doi.org/10.1037/a0032426

Bailey, K., Amlung, M. T., Morris, D. H., Price, M. H., Von Gunten, C., McCarthy, D. M., & Bartholow, B. D. (2016). Separate and joint effects of alcohol and caffeine on conflict monitoring and adaptation. *Psychopharmacology*, *233*(7), 1245–1255. https://doi.org/10.1007/s00213-016-4208-y

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Using **lme4**. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Braem, S., Bugg, J. M., Schmidt, J. R., Crump, M. J. C., Weissman, D. H., Notebaert, W., & Egner, T. (2019). Measuring Adaptive Control in Conflict Tasks. *Trends in Cognitive Sciences*, *23*(9), 769–783. https://doi.org/10.1016/j.tics.2019.07.002

Bugg, J. M., & Crump, M. J. C. (2012). In Support of a Distinction between Voluntary and Stimulus-Driven Control: A Review of the Literature on Proportion Congruent Effects. *Frontiers in Psychology*, *3*, 367. https://doi.org/10.3389/fpsyg.2012.00367

Cochrane, A. (2020). *TEfits: R package for streamlined nonlinear regression*. Zenodo. https://doi.org/10.5281/ZENODO.3788094

Cochrane, A., Simmering, V. R., & Green, C. S. (2019). Fluid intelligence is related to capacity in memory as well as attention: Evidence from middle childhood and adulthood. *PLOS ONE*, *14*(8), e0221353. https://doi.org/10.1371/journal.pone.0221353

Davelaar, E. J., & Stevens, J. (2009). Sequential dependencies in the Eriksen flanker task: A direct comparison of two competing accounts. *Psychonomic Bulletin & Review*, *16*(1), 121–126. https://doi.org/10.3758/PBR.16.1.121

Diamond, A. (2013). Executive functions. *Annual Review of Psychology*, *64*, 135–168. https://doi.org/10.1146/annurev-psych-113011-143750

Duthoo, W., Abrahamse, E. L., Braem, S., Boehler, C. N., & Notebaert, W. (2014). The Congruency Sequence Effect 3.0: A Critical Test of Conflict Adaptation. *PLoS ONE*, *9*(10), e110462. https://doi.org/10.1371/journal.pone.0110462

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & Psychophysics*, *16*, 143–149.

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition*, *2*, 101–118.

Fan, J., McCandliss, B. D., Sommer, T., Raz, A., & Posner, M. I. (2002). Testing the efficiency and independence of attentional networks. *Journal of Cognitive Neuroscience*, *14*(3), 340–347. https://doi.org/10.1162/089892902317361886

Gratton, G., Coles, M. G., & Donchin, E. (1992). Optimizing the use of information: Strategic control of activation of responses. *Journal of Experimental Psychology. General*, *121*(4), 480–506.

Halekoh, U., & Højsgaard, S. (2014). A Kenward-Roger Approximation and Parametric Bootstrap Methods for Tests in Linear Mixed Models—The *R* Package **pbkrtest**. *Journal of Statistical Software*, *59*(9). https://doi.org/10.18637/jss.v059.i09

Hutchison, K. A. (2011). The interactive effects of listwide control, item-based control, and working memory capacity on Stroop performance. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*(4), 851–860. https://doi.org/10.1037/a0023437

Kattner, F., Cochrane, A., Cox, C. R., Gorman, T. E., & Green, C. S. (2017). Perceptual Learning Generalization from Sequential Perceptual Training as a Change in Learning Rate. *Current Biology*, *27*(6), 840–846. https://doi.org/10.1016/j.cub.2017.01.046

Lehle, C., & Hübner, R. (2008). On-the-fly adaptation of selectivity in the flanker task. *Psychonomic Bulletin & Review*, *15*(4), 814–818.

Logan, G. D., & Zbrodoff, N. J. (1979). When it helps to be misled: Facilitative effects of increasing the frequency of conflicting stimuli in a Stroop-like task. *Memory & Cognition*, *7*(3), 166–174. https://doi.org/10.3758/BF03197535

Miyake, A. (2000). The Unity and Diversity of Executive Functions and Their Contributions to Complex "Frontal Lobe" Tasks: A Latent Variable Analysis. *Cognitive Psychology*, *41*(1), 49–100. https://doi.org/10.1006/cogp.1999.0734

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1–51). Lawrence Erlbaum.

Petersen, S. E., & Posner, M. I. (2012). The Attention System of the Human Brain: 20 Years After. *Annual Review of Neuroscience*, *35*(1), 73–89. https://doi.org/10.1146/annurev-neuro-062111-150525

Posner, M. I., & Petersen, S. E. (1990). The attention system of the human brain. *Annual Review of Neuroscience*, *13*, 25–42.

Rueda, M. R., Rothbart, M. K., McCandliss, B. D., Saccomanno, L., & Posner, M. I. (2005). Training, maturation, and genetic influences on the development of executive attention. *Proceedings of the National Academy of Sciences*, *102*(41), 14931–14936. https://doi.org/10.1073/pnas.0506897102

Schmidt, J. R. (2016). Proportion congruency and practice: A contingency learning account of asymmetric list shifting effects. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *42*(9), 1496–1505. https://doi.org/10.1037/xlm0000254

Schmidt, J. R., & Weissman, D. H. (2014). Congruency Sequence Effects without Feature Integration or Contingency Learning Confounds. *PLoS ONE*, *9*(7), e102337. https://doi.org/10.1371/journal.pone.0102337

Sidarus, N., Palminteri, S., & Chambon, V. (2019). Cost-benefit trade-offs in decision-making and learning. *PLoS Computational Biology*, *15*(9), e1007326. https://doi.org/10.1371/journal.pcbi.1007326

Spinelli, G., Perry, J. R., & Lupker, S. J. (2019). Adaptation to conflict frequency without contingency and temporal learning: Evidence from the picture–word interference task. *Journal of Experimental Psychology: Human Perception and Performance*, *45*(8), 995–1014. https://doi.org/10.1037/xhp0000656

Wendt, M., Kluwe, R. H., & Vietze, I. (2008). Location-specific versus hemisphere-specific adaptation of processing selectivity. *Psychonomic Bulletin & Review*, *15*(1), 135–140. https://doi.org/10.3758/PBR.15.1.135

Wenke, D., De Houwer, J., De Winne, J., & Liefooghe, B. (2015). Learning through instructions vs. learning through practice: Flanker congruency effects from instructed and applied S-R mappings. *Psychological Research*, *79*(6), 899–912. https://doi.org/10.1007/s00426-014-0621-1

Westlye, L. T., Grydeland, H., Walhovd, K. B., & Fjell, A. M. (2011). Associations between Regional Cortical Thickness and Attentional Networks as Measured by the Attention Network Test. *Cerebral Cortex*, *21*(2), 345–356. https://doi.org/10.1093/cercor/bhq101

Wühr, P., Duthoo, W., & Notebaert, W. (2015). Generalizing attentional control across dimensions and tasks: Evidence from transfer of proportion-congruent effects. *Quarterly Journal of Experimental Psychology*, *68*(4), 779–801. https://doi.org/10.1080/17470218.2014.966729

Supplemental Information

Supplemental Figures

Figure S1. Examples of stimuli. Any given trial would only have one row of stimuli. Any given block would only have one color of stimuli.

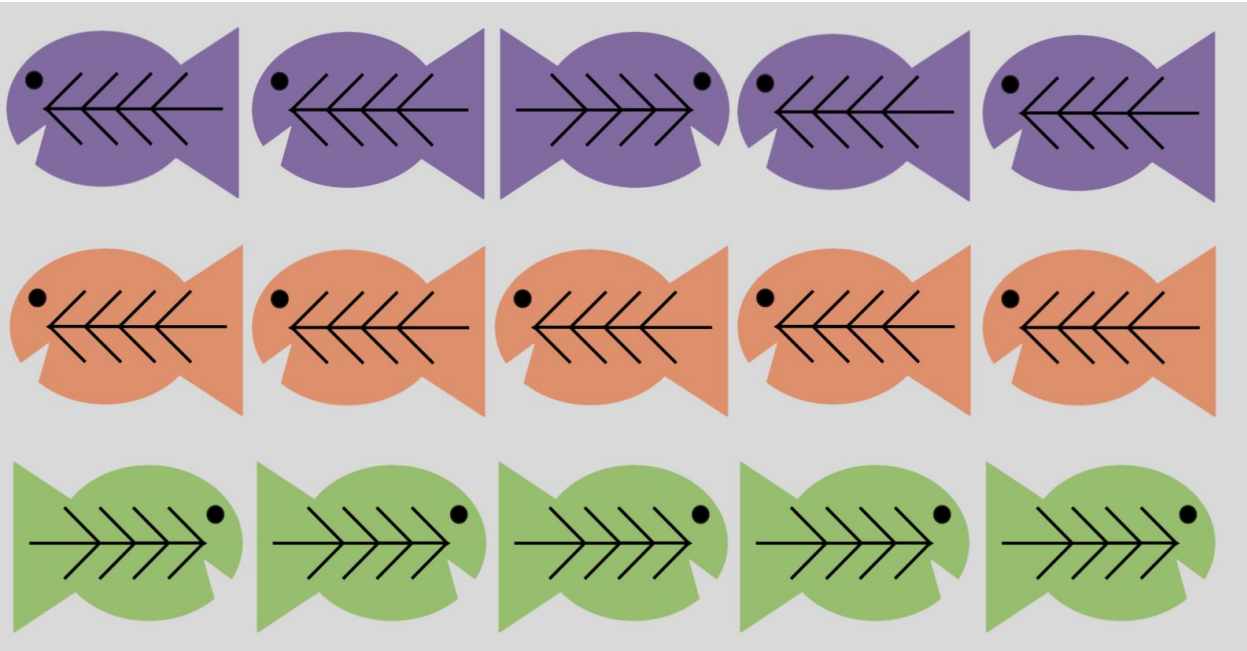
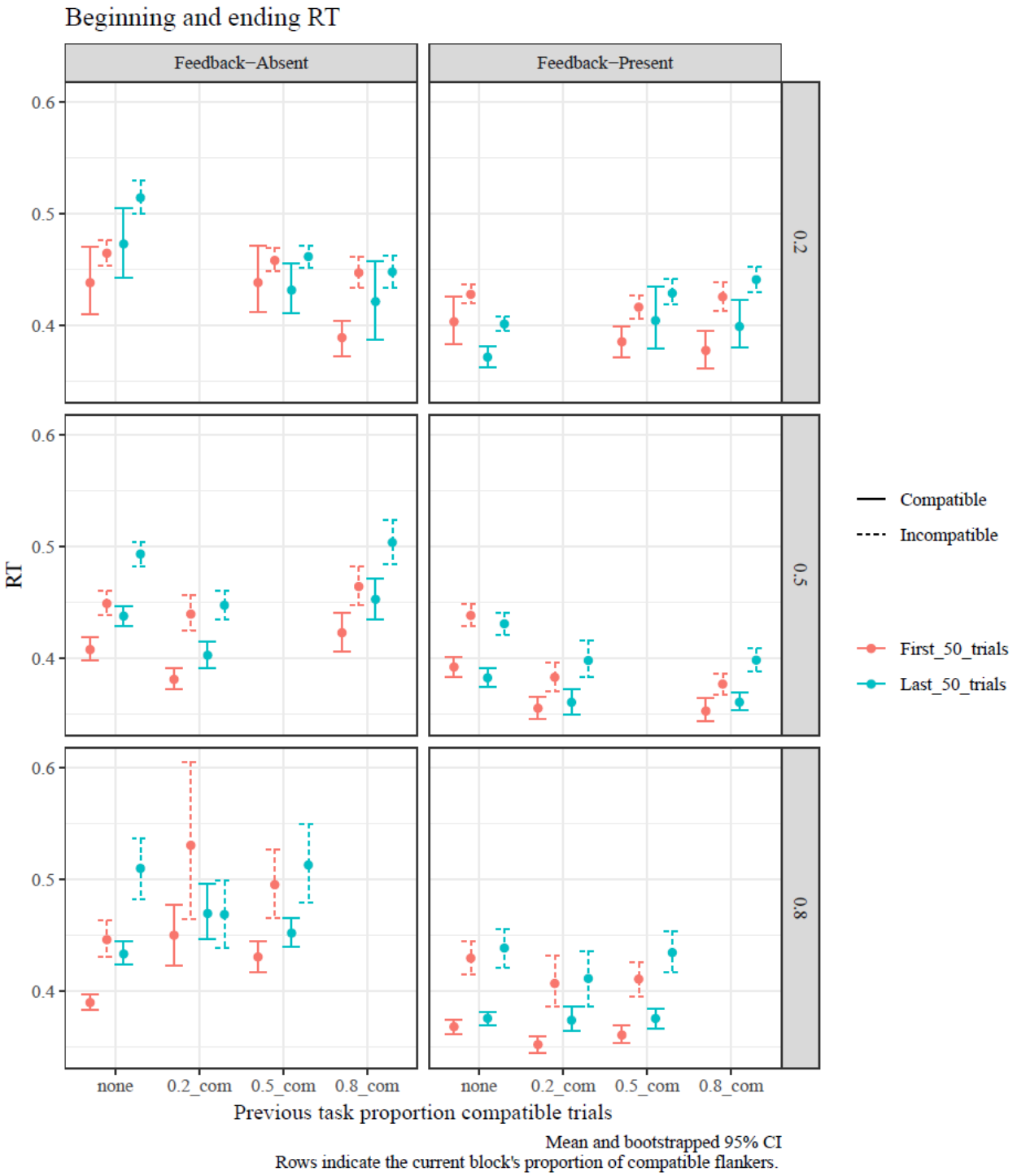


Figure S2. Results of both experiments, broken down by the first 50 and last 50 trials. Between-subjects' means and 95% CI are indicated.



Supplemental Methods

By-trial estimates of flanker effect were fit to each block of each participant's data using the R package TEfits (*github.com/akcochrane/TEfits*). Models were parametrized as 3-parameter exponential change in an ex-Gaussian distribution of response times using the *errFun=exGauss_tau* option.

Compatibility-dependent coefficients were estimated for each of the exponential parameters (i.e., start, rate, and asymptote), allowing for each component of the time-evolving curve to be influenced by response compatibility effects.

Appendix 7. IAT supplemental information.

Model formulas

Study 4.3b:

```
latency ~ exp(noiseMean) + exp(rtAsym) + (exp(rtStart) - exp(rtAsym)) * (2^((1 - totalTrialNum)/(2 + 2^rtRate)))
        noiseMean ~ (is_word || subject)
        sigma ~ (1 || subject)
        beta ~ exp(rtAsym) + (exp(rtStart) - exp(rtAsym)) * (2^((1 - totalTrialNum)/(2 + 2^rtRate)))
        rtStart ~ incFirst * inc + (inc || subject)
        rtAsym ~ incFirst * inc + (inc || subject)
        rtRate ~ incFirst * inc + (inc || subject)
```

In which `inc` represents trial type (response compatibility) and `incFirst` indicates order of presentation.

Study 4.3c:

```
latency ~ exp(noiseMean) + exp(rtAsym) + (exp(rtStart) - exp(rtAsym)) * (2^((1 - trialNum)/(2 + 2^rtRate)))
noiseMean ~ (isWord || subID)
sigma ~ (1 || subID)
beta ~ exp(rtAsym) + (exp(rtStart) - exp(rtAsym)) * (2^((1 - trialNum)/(2 + 2^rtRate)))
rtStart ~ expGroup * isInc + (expGroup * isInc || subID)
rtAsym ~ expGroup * isInc + (expGroup * isInc || subID)
rtRate ~ expGroup * isInc + (expGroup * isInc || subID)
```

In which `isInc` represents trial type (response compatibility) and `expGroup` represents the experiment type (e.g., Good/Bad, Competent/Incompetent).

Analyses using Composite Bias scores

This section includes additional analyses of data from Study 4.3c. One important aspect of the study was that many measures were taken to assess bias in various ways. In the original study Cox (2015; Experiment 3) averaged these scores to develop an aggregate individual-level bias score. Instead of an average, which may be influenced by distributional irregularities between measures, I re-calculated individual-level bias scores using a nonparametric method. I calculated nonmetric [isotonic] one-dimensional scaling of the behavioral measures in order to extract a latent prejudicial-behavior variable while avoiding possible violations of parametric assumptions. These 1-dimensional scales correlated

highly with the first PCA components of the variables (race r = 0.89; sex r = 0.89) indicating that the 1-dimensional scaling captured essentially the same variance as a dimension parametrically extracted from the behavioral data.In this Appendix I report Sex Composite and Race Composite bias scores.

Note that the critical |*r*| values would be .221 for the alternative model and .135 for the null model (given a sample size of 143 and a minimum raw Bayes factor of 3, using defaults of the R package **BayesFactor**).

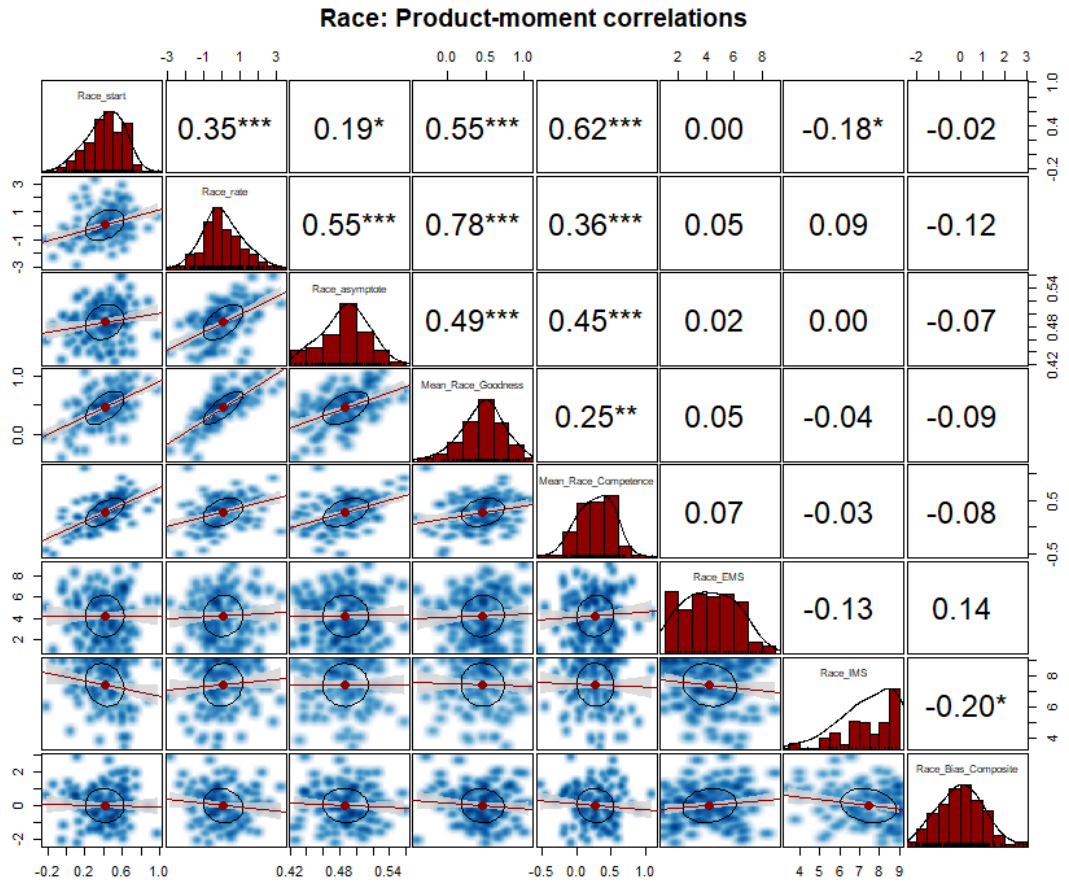


Figure A1. Composite bias score is far right column. No IAT scores are related to this composite. However, it is interesting that the only relation between race IAT and IMS or EMS was that starting bias levels shared about 3% of their variance with IMS

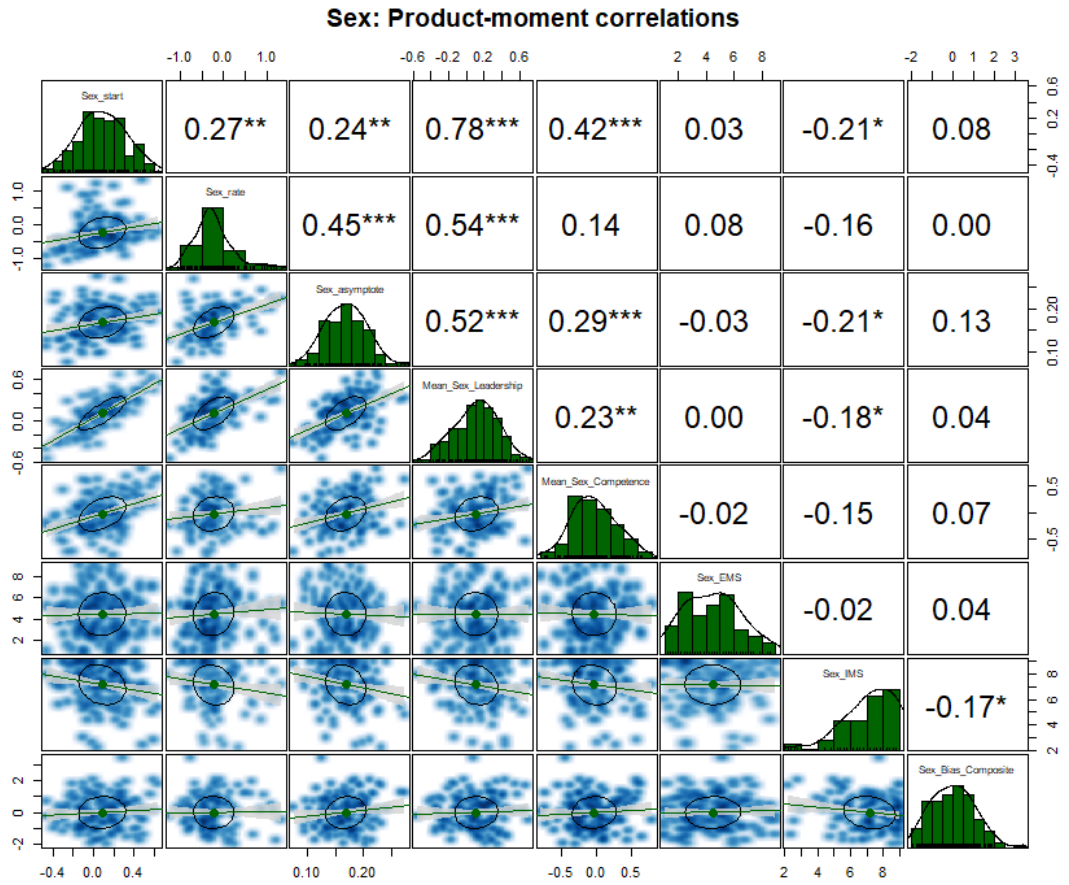


Figure A2. Composite bias score is far right column. No IAT scores are related to this composite. Various measures are related to IMS.

Appendix 8. Dale, Cochrane, & Green (under review)

Individual Difference Predictors of Learning and Generalization in Perceptual Learning

Gillian Dale[1], Aaron Cochrane[2], and C. Shawn Green[2]

[1] Environmental Sustainability Research Centre, Brock University, Canada

[2] Department of Psychology, University of Wisconsin-Madison, USA

Author Note

We have no conflicts of interest to disclose.

Correspondence concerning this article should be addressed to: Gillian Dale, Brock University, 1812 Sir Isaac Brock Way, St. Catharines, ON, Canada, L2S 3A1. Email: gdale@brocku.ca

Abstract

Given appropriate training, human observers typically demonstrate clear improvements in performance on perceptual tasks. However, the benefits of training frequently fail to generalize to other tasks, even those that appear similar to the trained task. A great deal of research has focused on training task characteristics that influence the extent to which learning generalizes. However, less is known about what might predict the considerable variations exist across individuals. Here we performed an individual differences study with the goal of identifying basic cognitive abilities and/or internal traits that are predictive of an individual's ability to learn and/or to generalize learning in perceptual learning tasks. We first showed that the rate of learning and the asymptotic level of performance that is achieved in two different perceptual learning tasks (motion direction and odd-ball texture detection) are correlated across individuals, as is the degree of immediate generalization that is observed and the rate at which a generalization task is learned. This indicates that there are indeed consistent individual differences in perceptual learning abilities. We then showed that several basic cognitive abilities and dispositional traits are associated with an individual's ability to learn (e.g., simple speed of processing; sensitivity to punishment) and/or generalize learning (e.g., cognitive flexibility; openness to new experiences) in perceptual learning tasks. Interestingly though, some abilities/traits that one might have a priori expected to be related to perceptual learning/generalization processes were not seen to be in our data set (e.g., fluid intelligence, grit/persistence, etc.). We suggest that the observed individual difference relations may suggest possible targets for future intervention studies meant to increase perceptual learning or generalization abilities.

Introduction

When human observers are given sustained training on a perceptual task, they typically demonstrate clear and sustained improvements in performance on the trained task itself (Dosher & Lu, 2017; Gibson & Gibson, 1955; Green et al., 2018; Maniglia & Seitz, 2018; Sagi, 2011; Seitz, 2017; Watanabe & Sasaki, 2015). For example, if participants are repeatedly shown two intervals of dot motion and are asked if the direction of motion in the two intervals was the same or was offset by 4°, participants show a clear increase in d' (sensitivity) over the course of training (e.g., Ball & Sekuler, 1982). Similarly, if participants are presented with a texture pattern composed of either all similarly oriented lines or the same basic pattern but where one of the lines is presented in a different orientation from the rest, participants typically show dramatic reductions in the minimum presentation duration that is required in order to differentiate these two types of patterns (Ahissar & Hochstein, 1997).

Although individuals usually do show improvements in performance on the very task on which they are trained, it is often the case that this training does not generalize to other tasks, even if they are highly similar to the training. For example, seemingly minor alterations to the size, direction, appearance, spatial or retinal location, or the orientation of stimuli can sometimes (but not always) result in performance dropping all the way back to pre-training levels (Ahissar, 1999; Ahissar & Hochstein, 1993; 1997; Ball & Sekuler, 1982; Fahle & Morgan, 1996; Fiorentini & Berardi, 1981; Poggio et al., 1991). This lack of generalization to even highly similar tasks, also known as *specificity*, is of clear theoretical interest, as it speaks to possible mechanisms underlying the perceptual learning process (Ahissar, Nahum, Nelken, & Hochstein, 2009; Lu & Dosher, 2009; Maniglia & Seitz, 2018; Shibata, Sagi, & Watanabe, 2014). It is also particularly problematic given the goal of applying perceptual learning to real-world situations

(Deveau & Seitz, 2014). Indeed, in order to have real-world value, it is typically necessary for the benefits of training to generalize across a variety of task parameters and situations. As such, a number of researchers have attempted to disentangle the many factors that might contribute to whether or not learning on a perceptual task is more generalizable or specific.

Training Task Factors Affecting Generalization

In examining factors that influence the degree of learning specificity that is observed, arguably the majority of work in the perceptual learning literature has focused on those factors inherent to the training tasks. For example, one factor that has been repeatedly shown to influence learning specificity is that of training task difficulty. In one seminal illustration of this effect, participants were trained on a texture discrimination task in which they had to determine whether one out of 49 lines in a rapidly presented 7 x 7 display was oriented differently than the other lines (Ahissar & Hochstein, 1997). Within this basic task, several aspects of the task that could affect difficulty were manipulated. These aspects included the possible location(s) the odd line could appear (more possible locations being harder than fewer possible locations) and the difference in orientation between the odd and distractor lines (smaller differences being harder than larger offsets). In all cases, the authors found that participants showed evidence of learning on the trained task itself (i.e., the participants were able to detect the presence of the oddball line with shorter presentation durations in both easier and harder task versions). Critically, though, when the authors assessed learning generalization by swapping the orientation of the background lines and the oddball lines, the authors found that on easier conditions learning generalized quite readily, but as task difficulty increased learning became much more specific (Ahissar, 1999; Ahissar & Hochstein, 1997). A similar result was shown by Liu and Weinshall (2000), who demonstrated that easy versions of a two-interval motion direction task led to significant

generalization (when the motion in the two intervals were either the same direction or were offset by 8°), while previous research had shown that more difficult versions led to significant specificity of learning (when the motion in the two intervals were either the same direction or were offset by 4°). It is worth noting though that while task difficulty is clearly an important determinant in the degree of observed generalization, some work has suggested that it is not the difficulty of the training task that governs the extent to which learning is specific or generalizes, but it is instead the difficulty of the transfer task that is the driving factor (i.e., there is more generalization to transfer tasks that require less precision; Jeter et al., 2009).

A second broad factor that appears to influence the degree of learning specificity that emerges from training is the degree of variety that is experienced (Deveau, Lovcik, & Seitz, 2014; Deveau & Seitz, 2014; Fulvio, Green, & Schrater, 2014). For example, several reports have now demonstrated that while learning tends to be specific to the trained retinal location, training participants on a different task at new locations – or even simply exposing participants to task-irrelevant stimuli in the new locations - will result in substantial generalization of the original trained task to the new locations (Wang et al., 2012; Xiao et al., 2008). Similar work has shown that presenting task-irrelevant stimuli that are oriented differently than the training stimuli can prevent adaptation to a specific orientation, thereby leading to greater generalization (Harris et al., 2012). Yet, while variety appears to be an important factor influencing learning generalization, there do appear to be clear boundary conditions. For example, in the case of passive exposure, generalization only appears to occur if the passive exposure to non-target locations is conducted simultaneously with the training task, as opposed to exposing the non-target locations prior to training (Zhang et al., 2010). Furthermore, the variety-linked generalization does not appear to occur if training is at threshold (Hung & Seitz, 2014).

A third factor that seems to play a role in determining the extent of learning generalization is the length of the training itself. In perhaps the clearest illustration of this, Jeter and colleagues had participants complete either one through six sessions (one session per day) of 1248 trials of a Gabor orientation discrimination task (Jeter et al., 2010), and examined how well the training generalized to a new version of the task. They found that the single session group demonstrated the highest amount of generalization, while the six-session group demonstrated the least, suggesting that greater amounts of practice on a given task will lead to more specificity.

And beyond the three broad factors above, there are a host of others that similarly appear to impact the observed degree of learning specificity. For example, training using longer staircases (and thus more prolonged training at or near threshold) has been seen to produce greater specificity of learning than training using shorter staircases (Hung & Seitz, 2014). Other factors that have been linked with the extent of perceptual learning and generalization include the amount of sleep between learning sessions (Karni et al., 1994), the noise present in the training stimuli (DeLoss, Watanabe, & Andersen, 2015), the amount of top-down attentional control required (Byers & Serences, 2012), the type of response that is required (Green et al., 2015), the concordance between training and transfer tasks (Snell et al., 2015), and the feedback given during training (Herzog & Fahle, 1997; Watanabe & Sasaki, 2015).

Individual Differences in Perceptual Learning/Generalization

Although it is increasingly well-established that a number of training task characteristics influence the extent to which learning is specific or more general, relatively less is known about what might predict the considerable variations in learning ability and generalization that exist within individuals. Indeed, although not typically a foci of the literature, it is clear that considerable individual differences in both the ability to acquire the trained task and in the ability

to generalize learning to new tasks do exist (Baldassarre et al., 2012; Fahle & Henke-Fahle, 1996; Green et al., 2015; Schmidt & Bjork, 1992; Withagen & van Wermeskerken, 2009). And while learning results in the field are, often for the sake of discourse, labeled according to the binary categories of "specific" or "generalized," in practice, like almost all areas of human performance, learning outcomes are usually much more continuous in nature. From both a practical and theoretical perspective, it is important to understand why these variations arise because understanding the factors that lead to more or less learning and generalization could help researchers develop more tailored training regimens, or even help with selection of viable candidates for learning studies.

There have been some recent attempts to understand how this inter-individual variability in perceptual learning/generalization arises. Some of these have examined how individuals use information from the tasks themselves, such as the variables that they focus on during learning, or how feedback influences their learning and generalization (Jacobs et al., 2000; Withagen & van Wermeskerken, 2009). Additionally, some current studies have identified neurophysiological differences that can predict learning and generalization, such as changes in alpha oscillations during learning (Freyer et al., 2013), and functional connectivity in visual areas and the prefrontal cortex (Baldassarre et al., 2012; Martin et al., 2012). However, somewhat less attention has been given to easily measured, naturally occurring, individual characteristics that may influence whether perceptual learning is specific or general.

Given that a more individual-differences based approach is somewhat rare in the domain of perceptual learning, it is worth considering other learning-based domains where this approach is more common such as in occupational and educational domains. In these, just as in the perceptual learning field, there is typically a large degree of individual variation in the ability to

learn a new task, skill, or process, as well as in the ability to generalize this new knowledge or ability to the actual workplace or classroom. Yet, unlike in the perceptual learning domain, there have been many studies over the years focusing on the individual factors that predict on-the-job and classroom learning, both to help develop better and more individualized training programs, but also to aid in personnel selection.

Personality, for instance, is one broad individual-difference domain that has been shown to be related to both learning and generalization in education and the workplace. For example, conscientiousness has been shown to positively correlate with GPA (Richardson & Abraham, 2009), and on-the-job learning and performance (Barrick & Mount, 1991; Blume et al., 2010; Burke & Hutchins, 2007; Schultz et al., 2011), as well as generalization of employee training to actual job performance (Blume et al., 2010; Colquitt et al., 2000). Extraversion and openness to experience were also found to be predictors of successful learning in some studies (Barrick & Mount, 1991), although not others (Blume et al., 2010; Schultz et al., 2011). Additionally, high levels of neuroticism were associated with greater generalization of training in the workplace (Blume et al., 2010). A number of other factors, including motivation to learn and attitude toward learning (Schultz et al., 2011), self-efficacy (Machin & Fogarty, 2012; Pugh & Bergin, 2006), and the relationship between the trainee and their instructor or manager (Grossman & Salas, 2011), also influence the degree of both learning, and generalization of learning, in workplace and educational settings. Beyond personality/dispositional factors, a host of general cognitive abilities have also been shown to be predictors of learning and generalization in the workplace and classroom. Such constructs include those meant to tap reasoning/fluid intelligence (Blume et al., 2010; Colquitt et al., 2000; Grossman & Salas, 2011), spatial reasoning (Uttal,

Miller, & Newcombe, 2013), working memory (Holmes & Gathercole, 2014), and executive functions (Titz & Karbach, 2014).

Current Study

It is clear that there are large inter-individual variations in both learning and generalization on perceptual tasks. While there have been recent attempts to understand how factors inherent to the training tasks themselves relate to the observed level of learning specificity, as well as some endeavors to examine possible underlying neurophysiological correlates of variation in learning and/or learning generalization, there is less understanding of how naturally occurring cognitive and dispositional differences are related to this inter-individual variation. While job-training/education and perceptual learning are obviously very different in numerous ways, it is reasonable to anticipate that the intrinsic factors that influence learning and generalization in one domain may also influence learning in another. As such, the purpose of the current study was to conduct an exploratory examination of naturally occurring individual characteristics and abilities that might predict learning and generalization on tasks of perceptual learning.

To progress in that endeavor though, it was first necessary to assess the degree to which learning/generalization was similar across multiple perceptual tasks. Indeed, one critical first step toward an understanding of individual differences in perceptual learning involves within-subjects comparisons of learning trajectories across tasks. To the extent that enhancement of perceptual abilities is systematically related to other factors, it should be the case that the components of improvement on perceptual training tasks are consistent. In other words, before addressing what factors influence “perceptual learning”, it is important to first demonstrate concordance across performance in tasks meant to tap “perceptual learning.” More specifically, it is necessary to

assess possible commonality in key parts of the perceptual learning process in initial perceptual abilities (i.e., how well do participants perform at a perceptual learning task right away), the time it takes to learn (i.e., learning rate), and/or the asymptotic levels of performance that are reached (Kattner, Cochrane, & Green, 2017). Decomposition of perceptual learning into these dimensions not only allows clarity in tests of individual differences across tasks, but also facilitates a more mechanistically-grounded account of cross-participant variation (Ackerman & Cianciolo, 2000; Kattner, Cochrane, Cox, et al., 2017).

To meet these various aims, participants were trained on two tasks that have previously been shown to result in significant learning effects over a relatively short span of time (1-3 hours of training), as well as to produce partial-to-full generalization: a Dot Motion discrimination task, and a Texture detection task. A number of cognitive abilities (e.g., selective attention, processing speed, memory, reasoning, etc.) and dispositional characteristics (e.g., personality, sensitivity to punishment/reward, affect, grit, etc.) that might be related to either learning ability or generalization were also assessed. Our goal was to determine whether any of these cognitive or dispositional characteristics would predict perceptual learning and/or generalization on our tasks.

Method

Participants

A total of 35 University of Wisconsin-Madison undergraduate students (23 females, 12 males) ranging in age from 18 to 33 years (M = 20.5, SD = 2.9) participated in this study. The participants were recruited via posted advertisements and received $60 for completing the study. Six participants were ultimately excluded for not completing the tasks as instructed (2), or for

having sufficiently poor performance that the key dependent measures could not be appropriately computed on at least 1 training task (4), leaving a total of 29 participants in the final analysis.

Study Overview

The study took place over the course of 4 sessions (90 minutes each; always on separate days) that were scheduled within no more than ten total days (see below for a full description of the tasks and design). On the first day, participants completed several computerized tasks examining basic cognitive abilities as well as a number of trait/personality questionnaires. On the second day, they completed three blocks of training on one of two possible perceptual learning tasks. On the third day, they completed a fourth training block on the same perceptual training task from the day before, a generalization block on that task, and then three blocks of training on the second perceptual learning task. Finally, on the fourth day they completed a fourth block of training on the second perceptual learning task as well as a generalization block on that task. They finished the study by completing a number of other questionnaires.

Apparatus

Across the four sessions of the study, participants completed both pen-and-paper questionnaires as well as computerized tasks. The computerized tasks were created and controlled using MATLAB and the Psychophysics Toolbox (PTB-3; Brainard, 1997; Kleiner, Brainard, & Pelli, 2007). All tasks were performed in a dimly lit testing room on a Dell OptiPlex 780 computer with a 23-inch flat screen monitor with an unrestrained viewing distance of approximately 60 cm. All responses were made via manual button press on the keyboard, or with the computer mouse. Participants received instructions prior to each task and completed practice trials under the supervision of the experimenter, after which they completed the remaining trials on their own.

Stimuli and Design

Perceptual Learning/Generalization Tasks

As noted above, participants completed training on two different perceptual learning tasks – a Dot Motion task and a Texture task. Immediately following training for each, they also had the degree of learning generalization assessed (generalization to an untrained direction/orientation). These two perceptual learning tasks were chosen because there was reason to suspect that participants would show (A) clear learning on the tasks over the period of time utilized in the current study (e.g., 800 trials), and (B) some degree of learning generalization (i.e., there was unlikely to be pure specificity or generalization of learning across all participants). This design would thus allow us to determine whether any aspects of the perceptual learning and/or generalization process were correlated across the tasks (e.g., if individuals who learned the dot motion task quickly also learned the texture task quickly; or if individuals who showed a great deal of learning generalization on the motion task also showed sizeable generalization in the texture task). If such correlations across tasks in particular aspects of perceptual learning task performance were observed, we could then examine the extent to which individual differences in basic cognitive abilities or personality traits were related to those aspects of perceptual learning/generalization.

Dot Motion Task. A grey circle with a radius of 5 degrees of visual angle was presented on a black background in the center of the screen. On each trial, 50 small black dots were presented for 100ms moving at a rate of 1 degree per second. The direction of motion for the dots on each trial was drawn from a uniform distribution between 105 and 165 degrees (in other words, the reference direction was 135 degrees, and on half of the trials the target dots moved in a more vertical/counterclockwise direction, and on the other half of trials the target dots moved

in a more horizontal/clockwise direction). Drawing stimuli from uniform distributions around a reference angle, rather than a more traditional method of constants approach, is a method utilized by our group previously both for reasons of measurement and for providing variety that might be useful for generalization (Green et al., 2015; Kattner, Cochrane, & Green, 2017; Kattner et al., 2017; Snell et al., 2015).

Immediately following the presentation of the target dots, mask dots were presented for 500ms. The mask dots each moved in a random direction at a speed of 1 degree per second. After the presentation of the mask, participants were asked to indicate whether the dots were moving more vertically (up arrow key) or horizontally (down arrow key; note that piloting suggested that the "up/down" explanation of the task response, rather than the "clockwise/counterclockwise" explanation produced better understanding and compliance with the task goals). After each response, participants received a feedback tone that informed them whether or not they were correct.

Participants first completed 5 practice trials in which the target dots were colored red and were presented for 500ms (in order to help distinguish stimulus dots from the mask dots), and then completed 3 training blocks of 200 trials each using the protocol described above. These three blocks were completed within a single session (see Procedure section below for details). In the following session, participants completed one final training block, and then completed a transfer block of 200 trials. The transfer block was identical to the training blocks with the exception that the direction of dot motion was now centered on 225 degrees (with motion directions uniformly sampled between 195 and 255 degrees; i.e., offset by 90 degrees), and no feedback was provided. The training and transfer tasks were counterbalanced across participants

such that half of the sample received training on the 135 angle stimuli, as described above, and half received training on the 225 angle stimuli.

Texture Task. This task was adapted from Ahissar and Hochstein (1997). At the beginning of each trial, an 800 Hz tone was presented for 500ms to alert the participant to the onset of the trial. The tone was followed by a variable delay of between 1000 and 2000ms, after which a 7 x 7 item matrix of black lines was presented in the center of the screen. On half of the trials, all 49 lines were presented at a 16° angle ("same" trials), whereas on the other half of trials one of the 49 lines was presented at a 36° angle ("different" trials). The stimulus matrix appeared on the screen for a variable stimulus-to-mask SOA (16, 30, 90, 120, 300, or 500ms), after which it was replaced with a mask that remained on the screen until the participant made a response. After stimulus presentation, participants were asked to indicate whether all of the lines were facing in the same direction (right arrow key) or if one was facing in a different direction than the others (left arrow key). Participants received feedback displayed on the screen ("Correct" or "Incorrect") following each trial.

Participants completed 6 practice trials at an SOA of 700ms to familiarize themselves with the task, and then completed 3 training blocks of 210 trials each using the protocol described above (i.e., 15 repetitions of each combination of SOA and same/different). This training phase was completed within a single session (see Procedure section below for details). In the following session, participants completed one final training block, and then completed a generalization block of 140 trials. The generalization block was identical to the training blocks with the exception that the lines were now oriented at a 106° angle, with the odd lines oriented at a 126° angle, and no feedback was provided. The training and generalization tasks were counterbalanced across participants such that half of the sample received training on the 16° and

36° angle stimuli, as described above, and half received training on the 106° and 126° angle stimuli. See the analytical section below for how learning and generalization were calculated.

Cognitive Predictor Battery

Given the reasonable paucity of work on individual level predictors of perceptual learning and/or perceptual learning generalization, the tasks utilized in our cognitive predictor battery were chosen as they represent a variety of constructs (e.g., speed of processing, working memory capacity, visual attention) that have frequently been implicated in individual learning or generalization differences in other domains (e.g., education or job-related performance).

Reaction Time Tasks (Speed of Processing). Three different reaction time tasks were employed – a simple go task (press a button as soon as a stimulus appears), a simple discrimination task (an arrow appears pointing left or right, press the corresponding arrow key), and a 3AFC discrimination task (one of three boxes lights up on the screen, press the corresponding button). While these three tasks differed slightly in terms of the complexity of the necessary response (one possible key press, two possible key presses, three possible key presses), all three involved quite simple visuo-motor transformations and thus were considered to be measures of speed of processing, which has been repeatedly implicated as a major factor in learning and learning generalization (Edwards et al., 2013; Green, Pouget, & Bavelier, 2010; Heppe et al., 2016; Ross et al., 2016; Schubert et al., 2015). Note that for all three RT tasks a recursive trial-by-trial outlier rejection procedure was performed in order to remove trials on which RT was excessively long (i.e., greater than 3 SDs from the mean) after which average RT across the remaining valid trials was used as the dependent measure (lower RTs = faster speed of processing).

Simple Reaction Time. Each trial began with an empty square appearing in the center of the screen, followed by an 800 Hz warning tone that sounded for 500ms. Following the tone, there was a variable wait time of between 1000 and 2000ms, after which the square turned black. Participants were required to press the spacebar as soon as the square changed to black. After they responded, their reaction time in milliseconds was presented in the center of the screen for 1000ms. If they responded too early, they were given a warning in the center of the screen ("You responded too soon! Wait until the square changes!"). Participants completed 6 practice trials, followed by 50 experimental trials.

Discrimination Reaction Time. Each trial began with an 800 Hz warning tone that sounded for 500ms, followed by a variable wait between 1000 and 2000ms. Then, a solid black arrow was presented in the center of the screen pointing to either the left or the right. The participants had to indicate as quickly as possible whether the arrow was pointing to the left or the right of the screen by pressing the corresponding arrow keys on the keyboard. At the end of each trial, participants received feedback on their reaction time if they answered correctly, or a warning if they responded incorrectly ("Incorrect. Make sure you indicate the correct direction") or if they responded too early. Participants completed 6 practice trials, followed by 60 experimental trials (30 left, 30 right).

3-AFC Reaction Time. Each trial began with an 800 HZ warning tone for 500ms and 3 empty squares were presented side by side in the center of the screen. After a variable wait of between 1000 and 2000ms, one of the 3 squares turned black. Participants had to indicate as quickly as possible whether the left (left arrow), middle (bottom arrow) or right (right arrow) box had illuminated. Participants received a warning if they responded too early. At the end of each trial, participants received feedback on their reaction time if they answered correctly, or a

warning if they responded incorrectly ("Incorrect. Make sure you indicate the correct direction") or too early. Participants completed 6 practice trials, followed by 60 experimental trials (20 left, 20 middle, 20 right).

Task-Switching (Cognitive Flexibility). Another widely observed cognitive factor related to learning and learning generalization is cognitive flexibility – in particular the ability to task-switch and/or multi-task (Glass, Maddox, & Love, 2013). In our task-switching measure, participants were asked to classify digits as higher/lower than 5 or odd/even depending on instructions presented on the screen. On each trial, either a blue or a yellow box was presented in the center of the screen (Rogers & Monsell, 1995). A single digit (1, 2, 3, 4, 6, 7, 8, or 9) was presented in the center of the box. The color of the box indicated to participants which task they should perform on the digit. When the box was blue, participants were to classify the digit as higher/lower than 5. When the box was yellow, the participant was to classify the digit as odd/even. To reduce memory load, the classification criteria for each trial was also presented beneath the box in large black text ("ODD/EVEN" or "HIGH/LOW"). Participants were asked to classify the digit according to the classification scheme as quickly and accurately as possible by pressing labelled keys on the keyboard (separate pairs of keys labeled "high" and "low" respectively, and "odd" and "even"). The stimulus remained on the screen until the participant responded, after which there was a 1000ms interval before the next trial. Critically, the classification scheme changed predictably every 2 trials, thus every other trial was a "switch" trial (i.e., participants went from classifying odd/even → high/low or vice versa). Participants completed 10 practice trials, followed by 400 trials (100 each of high/low/odd/even). Performance was measured by examining the average reaction time (for correct trials only) for switch and non-switch trials separately. Additionally, switch costs were calculated as the

difference in RT for non-switch and switch trials (smaller switch costs/faster RTs = better cognitive flexibility).

Filtering (Selective Attention). Like cognitive flexibility, selective attention has also been commonly observed to relate to learning abilities and, as such, is a frequent target for cognitive training (Bavelier, Bediou, & Green, 2018; Edwards et al., 2013). The selective attention task that we employed was adapted from Ophir, Nass, and Wagner (2009). Each trial began with an 800 Hz warning tone that sounded for 500ms. Following the warning tone there was a variable delay of 1000 to 2000ms, after which a display of colored lines was presented in the center of the screen. Every trial contained 6 red target lines, and either 2 or 10 blue, green, or yellow distractor lines. The lines were randomly oriented to face either vertically, horizontally, or diagonally. Participants were instructed to attend to the orientation of the red lines only, while ignoring the distractor lines. The display remained on the screen for 100ms and then, following a 900ms delay, a second display appeared. On half of the trials the second display was identical to the first, and on half of the trials one of the six target lines had changed orientation (the distractor lines were always identical from the first to the second display). Participants were asked to indicate whether any of the red lines had changed orientation from the first to the second display by pressing either the left (no change) or right (change) arrow keys. Participants completed 4 practice trials, followed by 92 test trials (26 repetitions of each combination of change/no change and distractor number). Performance was measured by calculating a sensitivity score (hits – false alarms) as a function of distractor set size (2 or 10) (less reduction in sensitivity with increasing distractors = better selective attention).

OSPAN (Working Memory). Working memory abilities have been consistently tied to learning outcomes, at least partially by virtue of the links between working memory and fluid

intelligence (Bergman Nutley & Söderqvist, 2017; Karbach & Unger, 2014). The working memory task employed here was adapted from Turner & Engle (1989). It required participants to remember a series of letters while simultaneously performing mathematical operations. Each trial began with a 3000 Hz warning tone, which was followed by a mathematical operation that was presented in the center of the screen (e.g., "4 ÷ 2 + 5 = 7"). The problem remained on the screen for 5 seconds, and participants were required to indicate whether the equation was true (left arrow key) or false (right arrow key) before the problem disappeared. A single letter drawn from all letters in the alphabet (except "I", "N", "O", "X", or "Y") was then presented in the center of the screen for 1000ms, and participants were instructed to remember the letter for a later serial recall task. The next operation was then presented immediately following the presentation of the letter. After a variable number of operations/letters had been presented (set size of 2, 4, or 6), participants were prompted to write down as many letters as they could remember from the set in the order in which they were presented. Participants completed 6 practice trials (3 each of set size 1 and 2), and then completed a total of 25 test trials (each set size presented 8 times, randomly intermixed). OSPAN score was calculated in two ways: a "harsh" measure of the total number of letters correctly recalled in order, and a "lenient" measure of the total number of letters recalled, regardless of order. The final OSPAN score was the average of the harsh and lenient measures.

Ravens Advanced Progressive Matrices (RAPM; Fluid Intelligence). Fluid intelligence is likely the most widely noted individual difference level predictor of learning in many domains, including in education (where the construct in many ways originated; e.g., Binet & Simon, 1916; Ritchie & Tucker-Drob, 2018; Rohde & Thompson, 2007). In this study, the odd numbered items from the RAPM task were used to measure fluid intelligence. On each trial,

participants were presented with a pattern that had one piece missing. They were required to select one of 8 options that they felt best matched the pattern by pressing the corresponding number key on the keyboard. The patterns remained on the screen until the participant made a response. There were 18 trials in total, and participants were given 20 minutes to complete the task. Performance was measured by taking the total number of items out of 18 that were correctly answered.

Painting (Complex Learning). The final predictor task was one in which participants were inherently asked to generalize complex perceptual experience to new stimuli, adapted from Kornell and Bjork (2008). In the first phase of this task, 6 paintings by each of 12 artists (72 in total; downloaded from http://sites.williams.edu/nk2/stimuli/) were serially presented in the center of the screen for 3000ms each. The last name of the artist was printed below each painting, and participants were instructed to learn the association between the painting style and the artist name. Immediately following the completion of the training phase, participants completed a distractor task in which they were required to count backward by 3s from 547 for 15 seconds. Following the distractor task, participants were serially presented with 48 new paintings (4 new paintings by each of the 12 artists) and were asked to select the name of the artist who had painted each painting with the mouse. Accuracy for correctly identifying the artist was used as an index of performance on the task.

Dispositional/Lifestyle Habits Predictor Battery

The goal of the full dispositional/lifestyle habits predictor battery was to capture key traits that previous research has indicated is predictive of learning outcomes – including those related to personality, sensitivity to reward & punishment, motivation and persistence, and use of modern

media (e.g., Barrick & Mount, 1991; Blume et al., 2010; Burke & Hutchins, 2007; Large et al., 2019; Richardson & Abraham, 2009; Schultz et al., 2011).

Personality/Dispositional Factors.

NEO-PI Big Five Questionnaire. This questionnaire, adapted from Costa and McCrae (1992), was designed to assess all five dimensions of the Big-5 personality model. There were 10 questions for each of the five personality factors (openness to experience, conscientiousness, extraversion, agreeableness, and neuroticism). Each question was answered using a 5-point Likert scale, based on the degree to which the participant felt each item in the questionnaire corresponded to their own behaviors (1 = very inaccurate; 5 = very accurate). The average rating for each of the 5 personality factors was calculated, with higher scores reflecting higher levels of a given personality trait.

BIS/BAS. This task was adapted from Carver and White (1994) and is designed to measure approach and avoidant behavioral tendencies. The scale contains 24 statements; 7 that assess behavioral inhibition (BIS), 13 that assess behavioral activation (divided into 3 subscales: drive, fun seeking, and reward responsiveness), and 4 filler questions. Participants were asked to rate how well each of the 24 statements applied to his/her current mood on a scale from 1 (very true for me) to 4 (very false for me). Half of the items were reverse scored, and the average rating for each of the 4 subscales was calculated (BIS, BAS-D, BAS-FS, and BAS-RR). Higher scores reflected higher levels of a given behavioral tendency.

Sensitivity to Reward/Sensitivity to Punishment Questionnaire (SPSRQ). In this questionnaire, adapted from Torrubia, Avila, Molto, and Caseras (2001), participants were presented with 48 statements designed to assess their sensitivity to reward and punishment. Participants used a 4-point scale (4 = very much yes; 1 = very much no) to rate the extent to

which they believed each statement corresponded to their own behavior. A "sensitivity to punishment" score and a "sensitivity to reward" score was calculated by taking the average rating for items that correspond to each of the scales. Higher scores indicated higher levels of sensitivity.

Global/Local. To measure individual differences in breadth, we used a global/local shape task adapted from Kimchi and Palmer (1982). Participants completed a booklet that contained 24 global/local shape triads. The shape triads consisted of three hierarchical shapes (i.e., 3-4 small, local squares or triangles that formed a larger, global square or triangle) arranged with a standard figure on top, and two comparison figures on the bottom. For each triad, participants were required to circle the comparison figure that they felt best matched the standard figure, as quickly as possible. Eight of the triads were "test triads" in which one of the comparison shapes matched the standard shape at the global level (the overall shape outline matched the standard), and one matched at the local level (the smaller shapes matched the standard). The other 16 triads were "filler triads", wherein only one of the two comparison shapes matched the standard. A global bias score was calculated by summing the total number of global options that were circled for the test triads only, resulting in a score from 0 to 8. A high score reflects a global bias, and a low score reflects a local bias (see Dale & Arnell, 2013).

Positive and Negative Affect Schedule (PANAS). To measure trait affect (positive & negative) we used a modified version of the PANAS (Watson, Clark & Tellegen, 1988). The questionnaire contained a list of 20 adjectives that describe various moods (e.g., happy, angry, bored). Ten of the adjectives were positive in valence, and 10 were negative in valence. Participants were required to rate the extent to which they generally experience each mood in their everyday life using a 9-point Likert scale (0 = not at all; 8 = very much). A positive (PA)

and negative (NA) affect score was calculated by averaging the ratings provided for the positive and negative items for a maximum PA and NA score of 5. Higher scores reflected higher levels of PA and NA.

Motivation/Persistence Measures.

Persistence. This scale, adapted from Ventura, Shute, and Zhao (2013) measures the ability to persevere even the face of great difficulty. The scale contained 4 statements, and participants were asked to rate how well each statement corresponded to their own behavior on a scale from 1(very inaccurate) to 5 (very accurate). An overall persistence score was calculated by summing the ratings for the 4 items, for a maximum persistence score of 16. Higher scores reflected greater levels of persistence.

Grit. Grit is defined as "perseverance and passion for long-term goals" (Duckworth, Peterson, Matthews, & Kelly, 2007). The grit scale used for the current study was adapted from Duckworth et al. (2007). Our version contained 8 statements, and participants were asked to rate how well each statement applied to them on a scale from 1(not at all like me) to 5 (very much like me). Ratings were averaged together for an overall grit score, with higher scores reflecting greater levels of perseverance.

Lifestyle/Habits Measures.

Media Multitasking Index (MMI). In this questionnaire adapted from Ophir et al. (2009), participants were presented with 12 different media forms (print media, television, computer-based video, music, non-music audio, computer/video games, phone/cell voice calls, instant messaging, text messaging, email, web-surfing, computer applications) and were first asked to estimate the amount of hours that they spend using each medium in an average week. Next, participants were given a matrix and asked to indicate the degree to which they use each of

12 primary mediums in conjunction with a secondary medium (e.g., "How often do you text message while watching tv?") using a 4-point scale (0 = never; 1 = a little of the time; 2 = some of the time; 3 = most of the time). An MMI score was calculated in accordance with Ophir et al. (2009), with higher scores reflecting greater media multitasking behaviors.

Video Game Experience. This scale asks participants to list their estimated expertise (from 1 to 7 with 7 indicating high expertise), number of hours per week, and number of weeks per year that they play video games from each of 7 gaming genres. This questionnaire was used to screen for video game experience; however, none of the participants were experienced gamers thus data from this questionnaire was not analyzed further.

Full Description of Task Order by Day

On **Day 1**, participants completed several computerized predictor tasks in the following order: Raven's Advanced Progressive Matrices (RAPM), three different reaction time (RT) tasks, task switching, filtering, and the Operation Span (OSPAN) task. If participants finished before the full 90 minutes were up, they were administered the dispositional questionnaires in the following order: Global/Local, NEO-PI, BIS/BAS, PANAS, Persistence, SPSRQ, and Grit. The MMI and Video Game Experience questionnaires were not administered until the very end of Day 4. **On Day 2**, participants first completed the Painting task, and then completed 3 training blocks of either the Dot Motion task or the Texture task (counterbalanced). They also completed any remaining questionnaires leftover from Day 1. **On Day 3**, participants completed a final training block for the task on which they had trained on Day 2 (either Dot Motion or the Texture task), and then completed the generalization block for that task. After completing the generalization task, participants were required to take a 5-minute break, and then completed 3 training blocks of either the Dot Motion or the Texture task, such that individuals who trained on

the Dot Motion task on Day 2 trained on the Texture task on Day 3, and vice versa. Finally, on **Day 4** participants completed a final training block of the task on which they had trained on Day 3, and then a generalization block for that task. Lastly, they completed the MMI and Video Game Experience questionnaires, were debriefed, and were compensated for their time.

Results

Analytical Methods for Perceptual Learning/Generalization Tasks

As in our previous work, performance in each of the perceptual learning tasks was fit as a continuous function of time (hierarchical time-evolving logistic regressions fit to the Dot Motion perceptual learning and generalization tasks; *model objects and data included in Supplementary Data*; hierarchical time-evolving Weibull [Quick psychometric function] regressions fit to the Texture learning and generalization tasks; Kattner, Cochrane, & Green, 2017; Kattner et al., 2017). This in turn allowed us to parameterize performance as a change in the threshold of the appropriate psychometric function. This change was calculated as: *asymptote+(start-asymptote)*exp((1-trialNumber)/(2+2^{rate}))*. Lower threshold values indicate better performance on both tasks, and lower rate parameters indicate faster learning (i.e., fewer trials required to achieve learning).

All correlations reported below are Spearman rank correlations that represent relationships across/within the learning tasks. The significance threshold (given *n* of 29 and alpha of .05) is approximately 0.356, and the corresponding *t* value is 2.052. All regressions used bootstrapped robust linear models using the R package **TEfits** (Cochrane, under review) with 20,000 resamples with replacement. Stars indicating reliability are applied if 0 falls outside the 95% quantiles of bootstrapped parameter values. Out-of-sample delta R-squared (ΔR^2_{oos}) is estimated as median reduction of out-of-sample prediction error when fitting models to all

predictors as compared to all predictors, except for the predictor of interest. This cross-validation was also fit 20,000 times, but on random subsamples of 80% of the data without replacement. Reduction in error was tested on the remaining random 20% of data.

Did We Observe Learning and Generalization in the Perceptual Learning Tasks?

Before examining detailed patterns of relationships between and within tasks, it is first critical to demonstrate that the tasks met the basic criteria laid out in the introduction – namely that we observed learning on the tasks over the period of training (~800 trials) and that we observed some intermediate degree of generalization. First, with respect to learning during initial training, asymptotic thresholds were reliably lower than starting thresholds in both texture (paired *t*-test m_{diff} = -0.60, CI = [-0.69,-0.51], $t(28)$ = -13.3, d_{Cohen} = -4.9) and dot-motion (paired *t*-test m_{diff} = -18.7, CI = [-26.1,-11.3], $t(28)$ = -5.2, d_{Cohen} = -1.9)

Second, with respect to generalization, we observed reliable increases in thresholds from the end of training to the start of generalization (*generalization cost;* Wilcoxon signed-rank test Z = 5.78), as well as reliable decreases in thresholds from the start of training to the start of generalization (*generalization benefit;* Wilcoxon signed-rank test Z = -4.16; see Figure 1a). Generalization learning happened in less time than did initial learning (i.e., smaller rate parameters; Wilcoxon signed-rank test Z = -2.13, see Figure 1b). In sum, these results indicate a nuanced pattern of partial generalization, providing further justification for tests of individual differences in generalization (see Figure 1c). The following analyses, testing generalization parameters while controlling for training parameters, capture the generalization benefits and changes in rate of learning from training to generalization. In addition, we also test generalization cost.

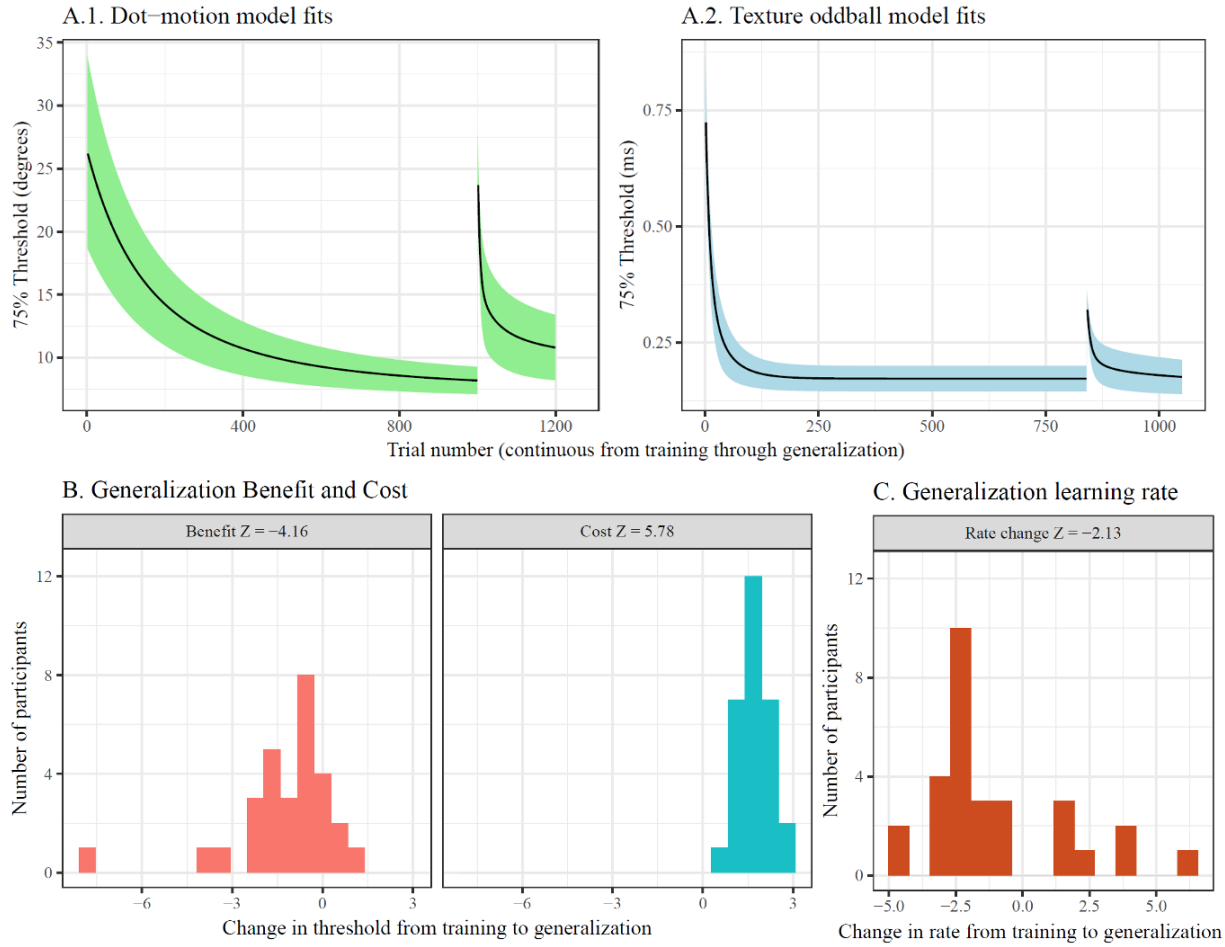


Figure 1. Panel A. *In both the dot motion task (**A.1**) and the texture task (**A.2**), participants showed clear evidence of learning during training. We also saw an intermediate degree of learning generalization (whereby initial performance on the generalization task was better than initial performance on the training task, but worse than asymptotic performance on the training task). This fact is plotted more explicitly in **Panel B** (left panel - values less than 0 mean better initial performance on the generalization task than on the training task; right panel – values greater than 0 mean that initial performance on the generalization task was worse than the asymptotic level of performance on the training task). Finally, as seen in **Panel C**, we also noted that participants learned the generalization task more rapidly than they had learned the training task (values less than zero indicating faster learning on the generalization task), suggesting that previous learning can speed the learning of subsequent tasks that share facets (i.e., learn to learn).*

Does Performance Correlate Across the Two Perceptual Learning Tasks?

First, we confirmed overall performance on the two perceptual learning tasks was related, with overall percent correct correlating at *Spearman* $\rho = 0.404$. As a blunt measure of perceptual-task association, we will use this as our baseline for comparing the more specific

relations between tasks. Note that, given our sample size, this reference correlation is associated with a two-tailed *p* value of 0.03. We then moved to examining correlations between specific aspects of the learning process (starting performance, rate, and asymptote).

Starting-performance parameters were not correlated across learning tasks (ρ = -0.005). In contrast, the correlation of rate of learning (ρ = 0.475) as well as asymptotic performance parameters did exceed that threshold (ρ = 0.488). This pattern is interesting for several reasons. First, individuals do not appear to show broad systematic patterns in their initial task performance across tasks (e.g., such that dispositional factors or life experiences would have led to immediate high performance or impairment on both tasks). Early performance instead appears to add noise to estimates of individual-level variation (i.e., over and above task-level variation). Second, parameters related to specific learning characteristics provide improved inference about individual-level variation. In particular, rate and asymptote of learning provide relatively precise interpretations when compared to a coarse aggregated measure such as overall accuracy. Not only are the parameters themselves more interpretable, but each of them shares more variance than the aggregated measure. Here it is critical to note that the increase in variance explained is not necessarily additive, yet neither are the correlations simply reflections of collinearity (i.e., within tasks' rate and asymptote parameters share less than 30% of their variance; texture ρ: 0.501; dot-motion ρ: 0.33).

We then turned to examining relations in generalization across tasks. Overall block-level generalization accuracies were correlated even more highly than training-task accuracies (ρ = 0.472). In contrast, generalization parameters were less correlated than parameters estimated from training data (start ρ: 0.37, rate ρ: 0.31, asymptote ρ: 0.07).

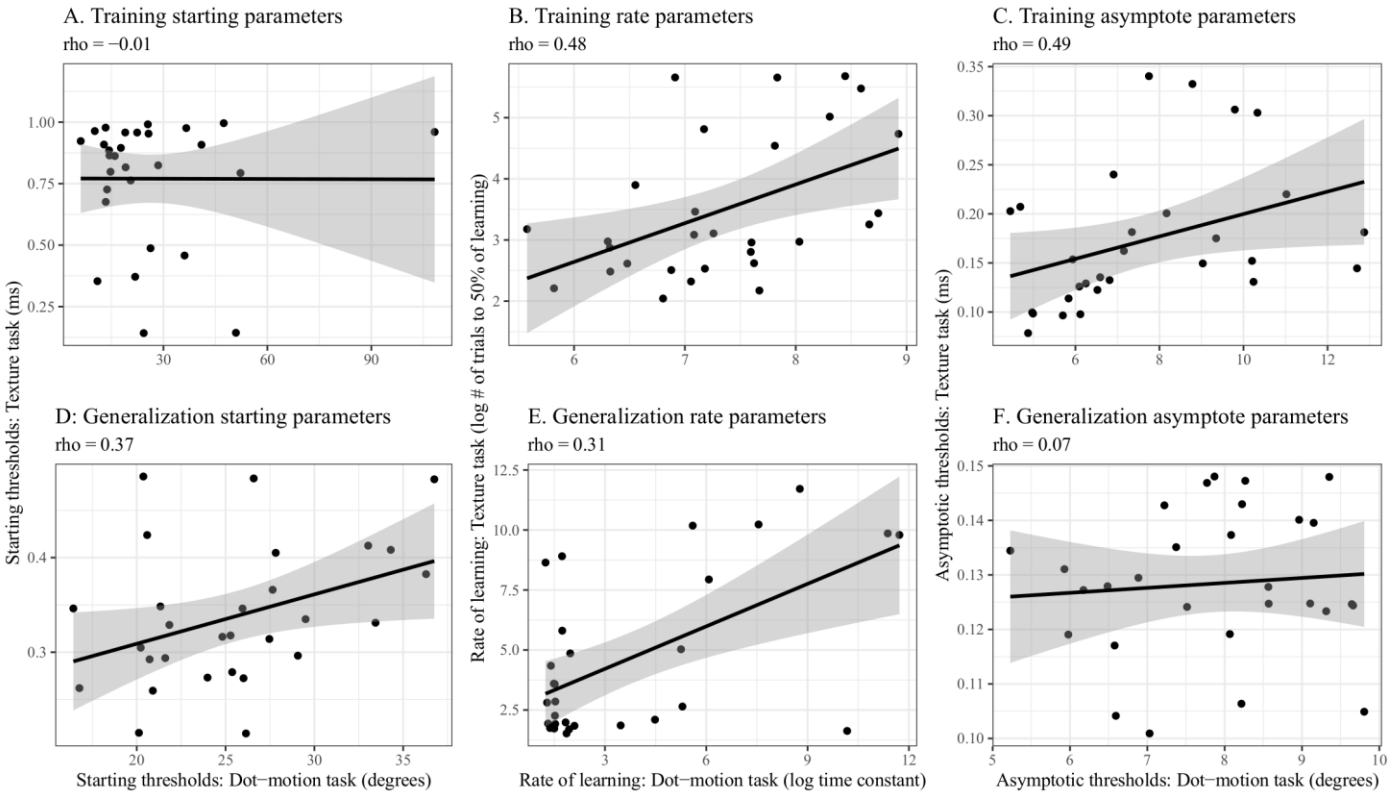


Figure 2. *Correlations between parameter estimates on Dot-motion perceptual learning (x axes) and Texture perceptual learning (y axes). The top row of plots shows parameter estimates during initial training while the bottom row shows parameter estimates during subsequent generalization. Left column shows starting thresholds, middle column is rate of learning, and right column shows asymptotic thresholds. Lines and shaded areas demonstrate a standard OLS best fit line and 95% CI. Rho is the Spearman rank-order correlation between tasks' parameter estimates. While most parameters show associations, training starting thresholds and generalization asymptotic thresholds do not.*

Is Generalization Related to Learning on the Tasks Themselves?

In order to test the individual-level relationships between training and generalization, geometric means were calculated across tasks for participants' parameters. This reduces the noise in the estimate of individual-level variation in learning starting points, rates, and asymptotes. The following analyses therefore consider each of the above three training parameters and three generalization parameters (i.e., starting point, rate, and asymptotic value).

A mixed pattern of correlations was found between training parameters and generalization learning parameters. Starting parameters in generalization and training did not meet our threshold for a reliable correlation (ρ = 0.383), with correlations between starting

parameters and other parameters being smaller still. In contrast, correlations between training and generalization asymptote and rate parameters were reliable (rate ρ = 0.503; asymptote ρ = 0.585). Notably, an even higher correlation was observed between training asymptote and generalization rate (ρ = 0.676), possibly indicating a mechanism of generalization in which rate of learning is particularly enhanced by initial learning (Kattner et al., 2017).

These findings provide further evidence that learning rate and asymptotic performance each reflect meaningful individual-level variation in generalization. There is also some evidence that the correlations are independent. Using robust regression to test the degree to which generalization rate is predicted by training asymptote and rate in a single model, only asymptote was found to be a reliable predictor (asymptote *b* = 3.00, CI = [0.39,9.20], ΔR^2_{oos} = 0.121; rate *b* = 0.55, CI = [-0.83,2.00], ΔR^2_{oos} = 0.023). In other words, training-task asymptotic performance is related to generalization learning rate even when controlling for training-task learning rate. The same applies to generalization asymptote, which is only related to training asymptote when the analogous model is fit (asymptote *b* = 0.17, CI = [0.06,0.31], ΔR^2_{oos} = 0.250; rate *b* = 0.015, CI = [-0.034,0.059], ΔR^2_{oos} = -0.058).

What Individual Difference Factors are Related to Perceptual Learning and Generalization?

Descriptive statistics of generalization measures are reported in Table S2. As discussed above, the three learning parameters (initial performance, rate, asymptote) reported here are composite scores formed by calculating the geometric means between the Dot Motion and the Texture task parameters. The following are the coefficients of various predictors in bootstrapped bivariate robust regression models with training-task parameters as outcome variables (see Table 1). Reported coefficient values are from models fit to the original non-resampled data. Lower

values are better in all cases (i.e., lower thresholds or times to learn). Caution should be taken when interpreting the baseline (pre-training) effects, however, due to the lack of correlation between the two tasks' starting thresholds.

Table 1. *Robust regression coefficients for the dispositional predictors of learning.*

Category	**Predictors**	**Baseline**	**Rate**	**Asymptote**
Cognitive	RT Composite	0.510**	0.420†	0.300†
	Task Switch Costs	0.030	0.074	0.270
	Filter Cost	-0.004	-0.180	-0.076
	OSPAN	-0.032	-0.027	-0.440*
	RAPM	0.040	-0.240	-0.180
	Painting	-0.300	-0.430*	-0.530*
Personality / Dispositional	Openness	-0.160	-0.096	0.015
	Conscientiousness	0.200	-0.046	-0.300
	Extraversion	0.055	-0.029	0.033
	Agreeableness	0.065	-0.100	-0.140
	Neuroticism	0.350*	0.460**	0.200
	BIS	0.380*	0.190	-0.140
	BAS-D	0.025	0.059	0.160
	BAS-FS	0.300*	0.370*	0.480*
	BAS-RR	0.270	0.350*	0.043
	Punishment Sensitivity	0.470*	0.310*	-0.140
	Reward Sensitivity	0.018	0.069	0.190
	Global Bias	-0.034	-0.360*	-0.190
	Positive Affect	-0.022	-0.037	0.018
	Negative Affect	-0.046	-0.034	-0.009
Motivation / Persistence	Persistence	-0.058	-0.220	-0.330
	Grit	0.300	0.025	-0.052
Lifestyle	MMI	-0.032	-0.027	-0.440†

*** indicates that 0 falls outside a parameter's bootstrapped 99% CI. * indicates that 0 falls outside a parameter's bootstrapped 95% CI. † indicates that 0 falls outside a parameter's bootstrapped 90% CI. All learning parameters and predictors are normalized (optimally Yeo-Johnson transformed then Z-scored) before model fitting to assist with interpretability.*

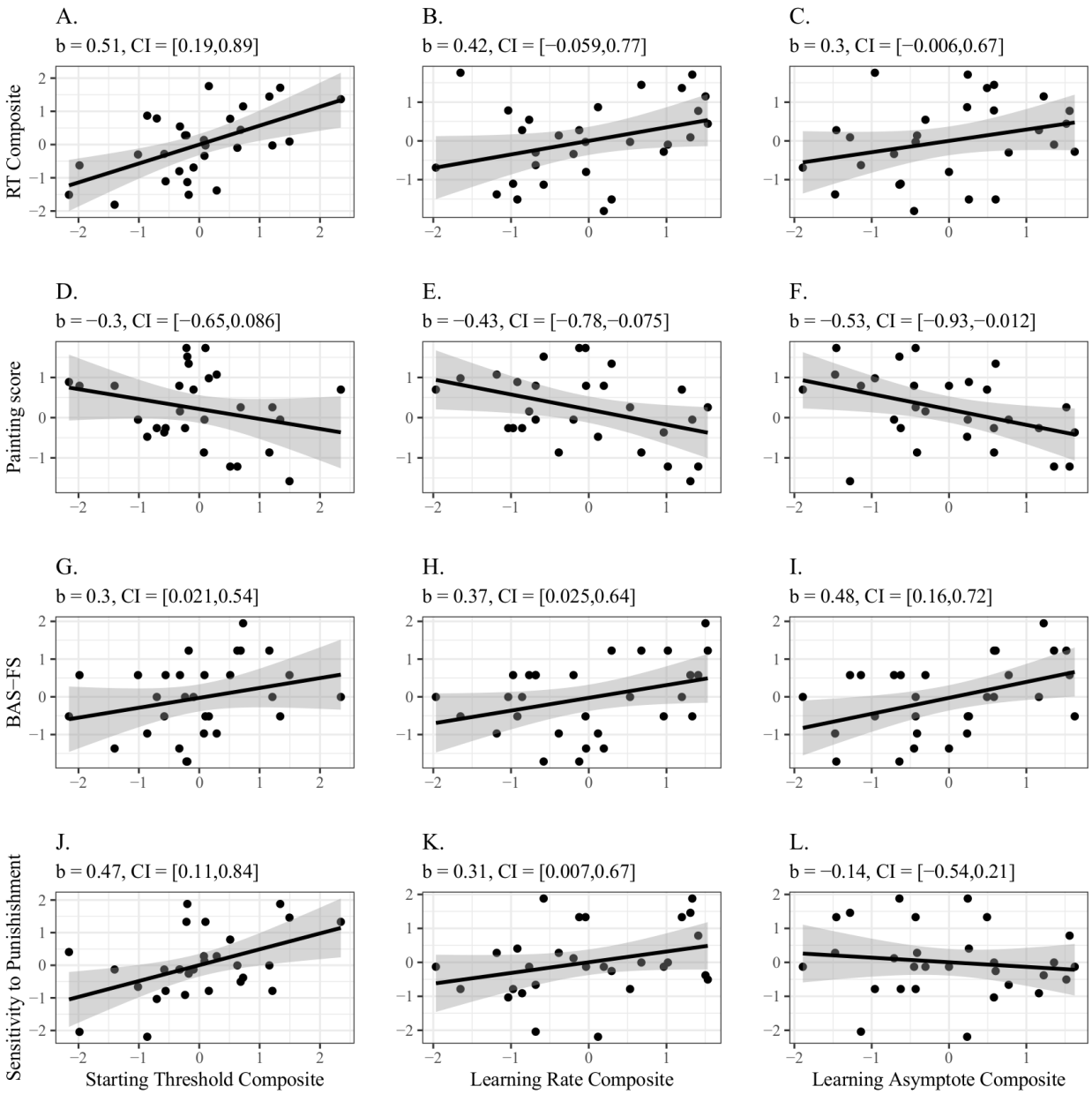


Figure 3. *Bivariate relations between four predictors (rows) and three components of learning in initial training (columns). Scores on the Painting learning task, BAS-FS, and Sensitivity to Punishment were each associated with multiple components of learning. Response Time (RT) Composite was associated only with variations in initial performance. While all statistics reported predict learning parameters, these parameters are on the x axis for the sake of plotting convenience. Scatterplots show Yeo-Johnson transformed variables. b and CI indicate the overall RLM slope and bootstrapped 95% CI of the slope. Lines and shaded areas are standard OLS linear regression fits and 95% CI.*

First, scores on the RT tasks measure predicted Pre-training (baseline) thresholds (*b* = 0.51, ΔR^2_{oos}= 0.289), such that lower RTs predicted lower initial thresholds. Additionally, lower

Neuroticism scores (b = 0.35, ΔR^2_{oos}= 0.150), lower BIS scores (b = 0.38, ΔR^2_{oos}= 0.145), higher BAS-FS scores (b = 0.30, ΔR^2_{oos}= 0.070), and lower Punishment Sensitivity scores (b = 0.47, ΔR^2_{oos}= 0.197) were each associated with lower initial thresholds. Next, higher scores on the painting task (b = -0.43, ΔR^2_{oos}= 0.193), higher global bias (b = -0.36, ΔR^2_{oos}= 0.109), lower neuroticism scores (b = 0.46, ΔR^2_{oos}= 0.181), lower BAS-FS (b = 0.37, ΔR^2_{oos}= 0.104), lower BAS-RR (b = 0.35, ΔR^2_{oos}= 0.105), and lower Punishment Sensitivity (b = 0.310, ΔR^2_{oos}= 0.083) were each associated with faster rates of learning. Finally, superior asymptotic performance (i.e., lower threshold) was predicted by higher scores on the OSPAN working memory task (b = -0.44, ΔR^2_{oos}= 0.144), higher scores on the painting category learning task (b = -.53, ΔR^2_{oos}= 0.252), and lower BAS-FS (b = 0.48, ΔR^2_{oos}= 0.177).

We next considered predictors of learning in generalization, above and beyond initial learning (see Table 2). Lower initial generalization threshold, controlling for initial training threshold (i.e., generalization benefit), was predicted by lower BAS-FS scores (b = 0.54, ΔR^2_{oos}= 0.244) and higher Persistence scores (b = -0.43, ΔR^2_{oos}= 0.064). Faster generalization learning rate, controlling for training learning rate, was predicted by smaller Task Switch Cost (b = 0.48, ΔR^2_{oos}= 0.207), lower openness to experience (b = 0.40, ΔR^2_{oos}= 0.193), and higher Punishment Sensitivity (b = -0.42, ΔR^2_{oos}= 0.222). Lower asymptotic threshold in generalization, when controlling for training asymptotic threshold, was predicted by higher Painting category learning scores (b = -0.48, ΔR^2_{oos}= 0.201) and by lower Grit scores (b = 0.29, ΔR^2_{oos}= 0.125). Finally, lower generalization cost, or initial generalization threshold controlling for final training threshold, was predicted by higher Neuroticism (b = 0.39, ΔR^2_{oos}= 0.110) and lower BAS-FS scores (b = 0.51, ΔR^2_{oos}= 0.209).

Table 2. *Robust regression coefficients for the cognitive and dispositional predictors of generalization. All parameters are controlled for training.*

Category	Predictors	Baseline	Rate	Asymptote	Cost
Cognitive	RT Composite	0.330	0.370†	0.300†	0.300
	Task Switch Costs	0.390†	0.480*	0.140	0.380†
	Filter Cost	0.120	0.110	-0.300†	0.150
	OSPAN	-0.220	-0.360†	-0.021	-0.120
	RAPM	-0.085	0.064	0.190	-0.055
	Painting	-0.150	-0.200	-0.480*	-0.110
Personality / Dispositional	Openness	0.220	0.400*	-0.170	0.190
	Conscientiousness	-0.350	-0.200	0.150	-0.240
	Extraversion	0.180	0.098	0.180	0.170
	Agreeableness	-0.270	-0.002	0.160	-0.290
	Neuroticism	0.340†	-0.056	-0.120	0.390*
	BIS	-0.120	-0.250†	0.004	0.038
	BAS-D	0.042	-0.110	-0.150	-0.001
	BAS-FS	0.540*	0.260	0.200	0.510*
	BAS-RR	0.150	0.008	0.042	0.230
	Punishment Sensitivity	-0.430†	-0.420**	-0.052	-0.190
	Reward Sensitivity	0.190	0.009	0.041	0.100
	Global Bias	-0.320	-0.100	0.095	-0.250
	Positive Affect	-0.016	-0.110	0.019	-0.007
	Negative Affect	0.100	-0.120	-0.037	0.130
Motivation / Persistence	Persistence	-0.430*	0.046	0.041	-0.410
	Grit	-0.270	0.067	0.290*	-0.084
Lifestyle	MMI	0.140	0.031	-0.063	0.230

*** indicates that 0 falls outside a parameter's bootstrapped 99% CI. * indicates that 0 falls outside a parameter's bootstrapped 95% CI. † indicates that 0 falls outside a parameter's bootstrapped 90% CI. All parameters and predictors are normalized (optimally Yeo-Johnson transformed then Z-scored) before model fitting to assist with interpretability. Cost refers to participants' starting generalization performance after controlling for asymptotic training performance.*

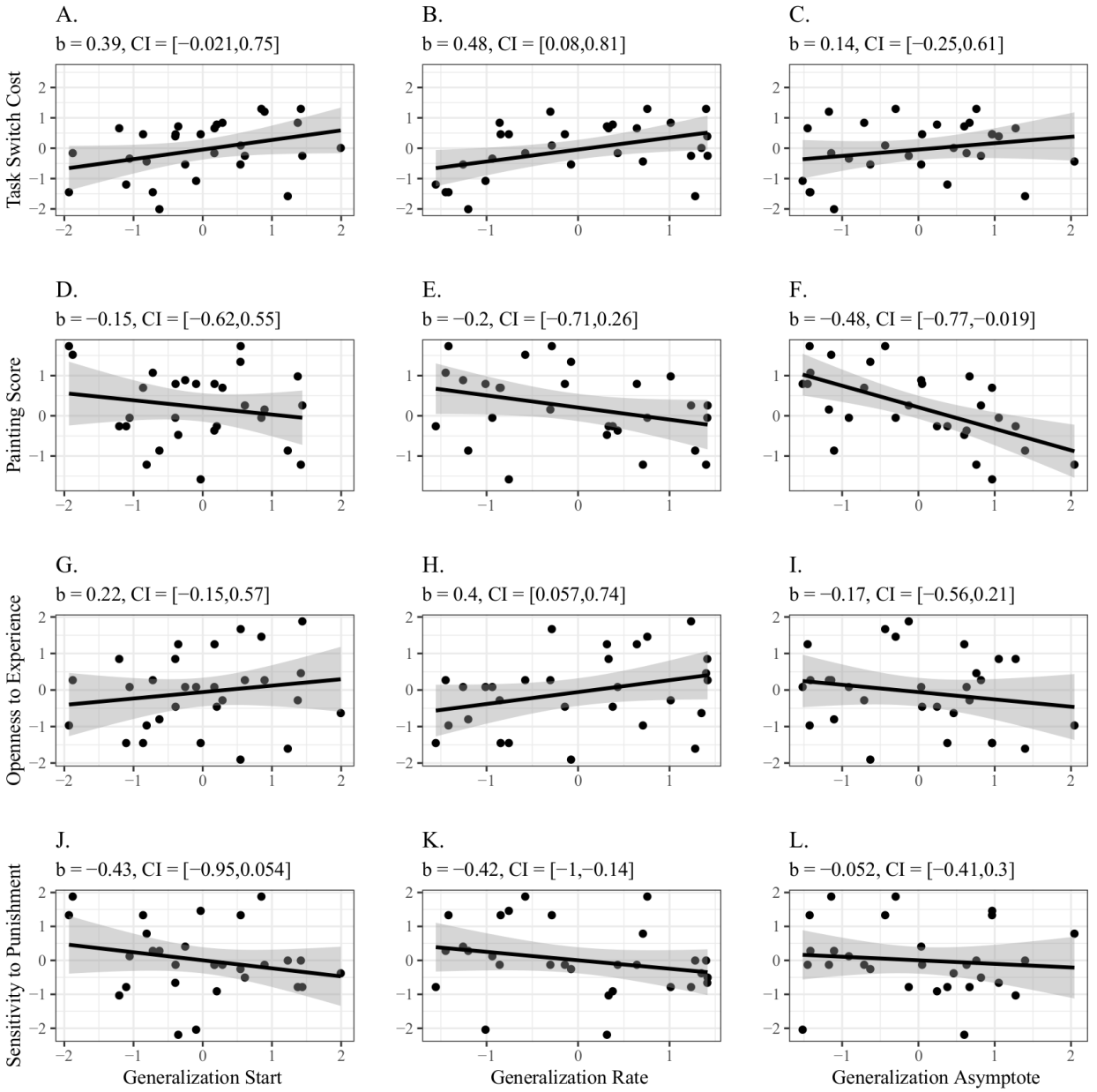


Figure 4. *Bivariate relations between four predictors (rows) and three components of learning in generalization (columns; these parameters controlled for the variance of the corresponding initial learning components). Rate of generalization is related to Task Switch Cost, Openness to Experience, and Sensitivity to Punishment. Asymptotic performance in generalization was related to scores on the Painting learning task. While all statistics reported predict learning parameters, these parameters are on the x axis for the sake of plotting convenience. Scatterplots show Yeo-Johnson transformed variables. b and CI indicate the overall RLM slope and bootstrapped 95% CI of the slope. Lines and shaded areas are standard OLS linear regression fits and 95% CI.*

Discussion

The purpose of the current study was to identify cognitive and dispositional predictors of learning and generalization of perceptual learning task performance. Although it is well documented that there are substantial inter-individual differences in learning and generalization on perceptual learning tasks, few studies have examined the individual characteristics that might predict which individuals show more or less learning or specificity of learning on these tasks. Given the recent interest in using perceptual tasks as a rehabilitation tool (e.g., to improve vision in individuals with amblyopia), and the larger literature on the beneficial effects of training on more complex tasks, such as video games (e.g., Bediou et al., 2018; Green & Bavelier, 2003; Powers et al., 2013; Toril, Reales, & Ballesteros, 2014; Wang et al., 2016), it is useful to attempt to identify easily-measured predictors of learning ability, and the propensity to generalize learning to a new task. By identifying such factors, it may be possible to use this information to tailor training regimens to individuals who fit a certain personality or cognitive profile, or identify ideal candidates for studies on the mechanics of learning and generalization. Ultimately, we were able to identify several dispositional and cognitive predictors of both learning and generalization on two perceptual tasks. In addition, we found a relationship between training asymptote and degree of generalization. These results are discussed in detail below.

Relationship between Learning and Generalization

The first critical finding of this study was that initial baseline performance on the Dot Motion task was unrelated to baseline performance on the Texture task, nor was baseline performance on either task related to generalization. As such, initial performance on these tasks was not predictive of whether or not participants would show generalization. However, both learning rate and asymptotic performance were significantly related across the two tasks. In

addition, both generalization rate and asymptote were associated with learning asymptote, such that individuals with a lower asymptotic threshold following training showed significantly better learning rates as well as lower asymptotic thresholds on the generalization task. This suggests that the individuals who came to the best level of performance at the end of the learning task (when controlling for baseline ability) had best encoded the underlying principles of the task, and were thus better able to apply these principles to the generalization orientation (i.e., learning to learn; Bavelier et al., 2012; Harlow, 1948; Kemp et al., 2010). This is consistent with previous work that has demonstrated a relationship between amount of learning and generalization (e.g., Duncan & Underwood, 1952; Lengyel & Fiser, 2019). Interestingly, learning asymptote predicted generalization asymptote over and above learning rate, suggesting that it is the degree of learning that explains the ability to generalize, rather than the speed at which an individual learns a given task.

Individual Difference Predictors of Learning

There was significant individual variation in both learning rate, and degree of learning, on our two tasks. Thus, our next goal was to determine whether cognitive and dispositional measures could predict this variation. Although our study was exploratory in nature, previous literature has demonstrated that cognitive factors, such as top-down attentional control (Byers & Serences, 2012) and matrix reasoning (Colquitt et al., 2000), are associated with increased learning. Additionally, several studies have suggested that various facets of personality, such as conscientiousness (Barrick & Mount, 1991; Blume et al., 2010; Burke & Hutchins, 2007; Richardson & Abraham, 2009; Schultz et al., 2011), extraversion (Barrick & Mount, 1991), openness to experience (Barrick & Mount, 1991), and neuroticism (Blume et al., 2010), are

associated with learning ability on a variety of tasks. As such, we anticipated that similar patterns would emerge in our study.

First, overall RT predicted initial thresholds on both tasks, such that individuals with faster RTs had significantly lower initial thresholds. In addition, individuals who scored lower on the "fun-seeking" aspect of the behavioural approach scale (BAS-FS) had significantly higher baseline thresholds. Finally, individuals with lower behavioural inhibition (BIS) scores, lower neuroticism, and lower sensitivity to punishment scores showed lower baseline thresholds.

Next, rate of learning was positively associated with performance on the painting task, such that individuals who were more accurate on the painting task showed faster rates of learning. In addition to the cognitive predictor of learning rate, we also found several dispositional predictors of learning. Individuals with a higher dispositional global bias (i.e., bias towards the wholistic, global letters rather than the individual, local letters on the global/local task) showed greater rates of learning. Furthermore, individuals who scored lower on both the "fun-seeking" and "reward responsiveness" facets of the behavioural approach scale (BAS) had significantly faster rates of learning. Lastly, individuals with lower neuroticism and lower sensitivity to punishment scores showed faster learning.

Finally, asymptotic performance was predicted by scores on both the OSPAN working memory task and the painting task, such that higher scores on both tasks was associated with lower asymptotic thresholds. "Fun-seeking" on the BAS scale was also associated with learning asymptote, such that individuals with lower BAS-FS scores had lower asymptotic thresholds (i.e., improved more on the task). The BAS scale measures several aspects of approach motivation, and is associated with reward/goal-seeking behaviours (both adaptive and maladaptive; Carver & White, 1994). It is unclear why BAS scores were associated with all three

measures of learning performance in our study, but perhaps the tasks themselves were not particularly rewarding, and thus individuals who were strongly driven by reward and/or punishment avoidance were not as motivated to learn on these tasks. Regardless, this suggests that further investigation is needed into the relationships between approach/avoidance motivation and learning.

Interestingly, many of the cognitive and dispositional measures that may have been expected a priori to relate to learning, based upon their relations in other learning studies, were not significant predictors here. For example, we anticipated that scores on the RAPM might relate to learning rate and asymptote, as demonstrated by Colquitt et al. (2000), but no such relationship was found. Additionally, although there was a relationship between baseline performance and RT, the relationship with learning rate and asymptote only approached, but did not reach significance. This is somewhat surprising as there is some suggestion that the link between cognition and perception is fundamentally reliant on an individual's ability to rapidly process information (e.g., Salthouse, 1993; 1996). For instance, Ackerman and Cianciolo (2000) found that correlations between psychomotor measures and a novel complex task increased across learning of the novel task, while correlations with knowledge or spatial measures decreased in the same time frame. While one possible explanation for this lack of effect is the sample size (i.e., that the relationship would have reached significance with a larger sample), if this is the case, it would still suggest that the relationship is not sufficiently strong to harness for any translational purposes. A second possible explanation is that our study used simple perceptual learning tasks, whereas other studies examined learning in more complex, real-world tasks and situations. As such, while the perceptual tasks used in this study are perceptually taxing, they are not necessarily cognitively taxing, thus possessing better working memory,

cognitive control, reasoning abilities, etc., may not necessarily be beneficial when performing these tasks.

In addition to the cognitive measures, we also expected that scores on several facets of personality would predict learning. In particular, conscientiousness (Barrick & Mount, 1991; Blume et al., 2010; Burke & Hutchins, 2007; Richardson & Abraham, 2009; Schultz et al., 2011), extraversion (Barrick & Mount, 1991), and openness to experience (Barrick & Mount, 1991) have previously been shown to relate to learning on a number of different tasks. In some ways, this may be framed as a positive for the field of perceptual learning. The use of college-student populations in this field, like many areas of psychology, has resulted in concerns regarding whether the findings will generalize to a less select population. In particular, students at high-ranking universities are somewhat, almost by definition, selected for conscientiousness. It is thus promising that this personality feature does not appear to play a major role in perceptual learning results. And although we did not replicate the expected links between personality and learning, we did show several relationships between neuroticism and learning, such that individuals with lower neuroticism scores showed lower baseline thresholds, and higher learning rates, than did individuals higher in neuroticism.

Predictors of Generalization

In addition to examining the factors that were associated with learning, we also were interested in identifying cognitive and dispositional predictors of learning generalization. Previous research has demonstrated that individual differences in matrix reasoning ability (Blume et al., 2010; Grossman & Salas, 2011), as well as the personality traits of conscientiousness (e.g., Blume et al., 2010; Colquitt et al., 2000), extraversion (Burke & Hutchins, 2007; Naquin & Holton, 2002), openness to experience (Burke & Hutchins, 2007;

Herold et al., 2002), and neuroticism (Blume et al., 2010) are also associated with an increased ability to generalize learning. As such, we anticipated that these same patterns might emerge in the current study.

First, baseline performance on the generalization task was associated with scores on the "fun-seeking" facet of the BAS scale, such that individuals with lower fun-seeking scores had lower thresholds on the task. This is congruent with the relationship between BAS-FS that was found for the learning task, such that lower approach motivation was associated with better performance. Additionally, persistence was associated with baseline performance, such that individuals with higher levels of persistence had lower initial thresholds.

Next, we examined the relationships between our cognitive and dispositional predictors and generalization rate. There was a significant relationship between rate of generalization and switch costs (i.e., the RT cost of changing from one task to another) on the task-switching measure, such that individuals with lower switch costs had a faster generalization rate. As such, some aspect of greater executive control may be associated with the ability to generalize learning to a novel task. In addition, we found a relationship between generalization rate and openness to experience, such that individuals who were less open to experience had faster generalization rates. This is in contrast to previous studies which have shown that greater openness to experience is associated with training proficiency in workplace settings, such that individuals higher in openness are better able to flexibly take on new roles and apply their acquired knowledge across several new domains (Burke & Hutchins, 2007; Herold et al., 2002). Lastly, sensitivity to punishment was related to generalization rate, such that individuals who were more sensitive to punishment had a faster rate of generalization. As such, the relationship between punishment/reward sensitivity and learning/generalization requires further investigation in order

to understand how dispositional levels of these traits affects both the ability to learn on a novel task, and to generalize that learning.

In addition to generalization rate, we also examined predictors of generalization asymptote. Higher scores on the painting task were associated with lower thresholds at asymptote (i.e., better generalization), which is congruent with our earlier finding that higher painting task scores were associated with lower learning asymptotes (i.e., better learning). Generalization asymptote was also predicted by scores on the grit measure, such that individuals lower in grit had lower asymptotic thresholds. Grit is defined as the ability to persist in the face of difficulty (Duckworth et al., 2007), thus it is unclear why "grittier" individuals would show less generalization on this task.

Finally, we examined the relationship between our predictors and generalization costs. Both scores on the "fun-seeking" facet of the BAS scale and neuroticism were significantly associated with generalization costs, such that individuals who scores lower on both of these measures showed fewer costs.

Overall, we found no relationships between matrix reasoning and generalization on our tasks. This is in contrast to the literature which suggests that greater reasoning and fluid intelligence skills are more likely to lead to generalization of learning (e.g., Colquitt et al., 2000). However, as mentioned previously, many of these studies used complex, real-world learning measures, thus it is possible that complex matrix reasoning is not required for the types of simple perceptual tasks utilized here. Again, as was true of the lack of correlations between learning/generalization and conscientiousness, the lack of correlations with matrix reasoning may be considered heartening in that the typical populations that have been studied in perceptual learning tasks have often been selected (e.g., via population bias) for higher matrix reasoning

abilities. Additionally, we found no relationships between generalization and extraversion. Previous research has shown that individuals high in extraversion are more motivated to improve their work by generalizing their knowledge to new situations (Burke & Hutchins, 2007; Naquin & Holton, 2002). Again, this null finding may be due to the type of task employed here.

Conclusions

This study was designed to identify possible cognitive and dispositional predictors of learning and generalization on perceptual tasks. First and foremost, we found a significant relationship between learning and generalization on our tasks. This is critical for many theories of perceptual learning, in that it suggests that there is a more global ability to learn to perform perceptual tasks and to reach high levels of performance (and that this isn't predicted by initial abilities on the tasks). This provides support for many ideas in the field that it may be possible to globally enhance the ability to learn perceptual tasks (which would not have been true had we found that performance on the tasks was totally independent). We were also able to identify several dispositional and cognitive factors that related to baseline ability, rate, and magnitude of learning and generalization. While both learning and generalization were primarily associated with factors relating to motivation and effort (e.g., BAS-FS, sensitivity to punishment, neuroticism), there were also several relationships between learning, generalization, and cognitive ability. For example, scores on the painting memory/learning task were associated with both learning and generalization asymptote, suggesting that some aspect of learning/memory/reasoning explains individual variation in perceptual learning. Interestingly, many variables that were expected to relate to learning and/or generalization, such as processing speed, matrix reasoning, reaction time, and some aspects of personality, did not emerge as significant predictors.

While there were several limitations to this study, including the purely correlational nature of the measures, these findings demonstrate, for the first time, that the large variations in learning and generalization that are generally seen on tasks of perceptual learning may not only be influenced by the types of tasks that are used, but also by naturally-occurring characteristics of the participant. Given the recent interest in using perceptual tasks for rehabilitation and general cognitive improvement, it is important to better understand how these individual characteristics influence learning and generalization. More research is necessary in order to better understand how the factors identified here influence the specificity of learning, and whether individuals who possess traits that are associated with better learning/generalization on the tasks will show enhanced learning/generalization across a broader spectrum of tasks.

Open Practices Statement

The data and materials for all experiments will be made freely available upon publication on an open-science platform. This experiment was not preregistered.

References

Ackerman, P. L., & Cianciolo, A. T. (2000). Cognitive, perceptual-speed, and psychomotor determinants of individual differences during skill acquisition. *Journal of Experimental Psychology: Applied*, *6*(4), 259–290. https://doi.org/10.1037/1076-898X.6.4.259

Ahissar, M. (1999). Perceptual learning. *Society*, *8*(4), 124–128. https://doi.org/10.1111/1467-8721.00029

Ahissar, M., & Hochstein, S. (1993). Attentional control of early perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*, *90*(12), 5718–5722. https://doi.org/10.1073/pnas.90.12.5718

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. *Nature*, *387*(6631), 401–406. https://doi.org/10.1038/387401a0

Ahissar, M., Nahum, M., Nelken, I., & Hochstein, S. (2009). Reverse hierarchies and sensory learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1515), 285-299. https://doi.org/10.1098/rstb.2008.0253

Baldassarre, A., Lewis, C. M., Committeri, G., Snyder, A. Z., Romani, G. L., & Corbetta, M. (2012). Individual variability in functional connectivity predicts performance of a perceptual task. *Proceedings of the National Academy of Sciences*, *109*(9), 3516–3521. https://doi.org/10.1073/pnas.1113148109

Ball, K., & Sekuler, R. (1982). A specific and enduring improvement in visual motion discrimination. *Science*, *218*(4573), 697-698. doi: 10.1126/science.7134968

Barrick, M. R., & Mount, M. K. (1991). The Big Five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1–26. doi: 10.1111/j.1744-6570.1991.tb00688.x

Bavelier, D., Bediou, B., & Green, C. S. (2018). Expertise and generalization: Lessons from action video games. *Current Opinion in Behavioral Sciences, 20*, 169-173. doi: 10.1016/j.cobeha.2018.01.012

Bavelier, D., Green, C. S., Pouget, A., & Schrater, P. (2012). Brain plasticity through the life span: Learning to learn and action video games. *Annual Review of Neuroscience, 35*, 391-416. doi: 10.1146/annurev-neuro-060909-152832

Bediou, B., Adams, D. M., Mayer, R. E., Tipton, E., Green, C. S., & Bavelier, D. (2018). Meta-analysis of action video game impact on perceptual, attentional, and cognitive skills. *Psychological Bulletin, 144*(1), 77-110. https://doi.org/10.1037/bul0000130

Bergman Nutley, S., & Söderqvist, S. (2017). How is working memory training likely to influence academic performance? Current evidence and methodological considerations. *Frontiers in Psychology, 8*, 69. https://doi.org/10.3389/fpsyg.2017.00069

Binet, A., & Simon, T. (1916). *The development of intelligence in children (The Binet-Simon Scale).* Williams & Wilkins Co: Baltimore.

Blume, B. D., Ford, J. K., Baldwin, T. T., & Huang, J. L. (2010). Transfer of training: A meta-analytic review. *Journal of Management, 36*(4), 1065–1105. https://doi.org/10.1177/0149206309352880

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433-436. doi: 10.1163/156856897X00357

Burke, L. A., & Hutchins, H. M. (2007). Training transfer: An integrative literature review. *Human Resource Development Review, 6*(3), 263–296. https://doi.org/10.1177/1534484307303035

Byers, A., & Serences, J. T. (2012). Exploring the relationship between perceptual learning and top-down attentional control. *Vision Research, 74,* 30–39. https://doi.org/10.1016/j.visres.2012.07.008

Carver, C. S., & White, T. L. (1994). Behavioral inhibition, behavioral activation, and affective responses to impending reward and punishment: the BIS/BAS scales. *Journal of Personality and Social Psychology, 67*(2), 319-334.

Cochrane, A. (under review). TEfits: Nonlinear regression for time-evolving indices. *Journal of Open Source Software*.

Colquitt, J. A., LePine, J. A., & Noe, R. A. (2000). Toward an integrative theory of training motivation: A meta-analytic path analysis of 20 years of research. *Journal of Applied Psychology, 85*(5), 678–707. https://doi.org/10.1037/0021-9010.85.5.678

Costa, P. T., & McCrae, R. R. (1992). Normal personality assessment in clinical practice: The NEO Personality Inventory. *Psychological Assessment, 4*(1), 5-13.

DeLoss, D. J., Watanabe, T., & Andersen, G. J. (2015). Improving vision among older adults: Behavioral training to improve sight. *Psychological Science, 26*(4), 456–466. doi: 10.1177/0956797614567510

Deveau, J., & Seitz, A.R. (2014). Applying perceptual learning to achieve practical changes in vision. *Frontiers in Psychology,* 5, 1166. doi: 10.3389/fpsyg.2014.01166

Deveau, J., Lovcik, G., & Seitz, A. R. (2014). Broad-based visual benefits from training with an integrated perceptual-learning video game. *Vision Research, 99,* 134–140. doi: 10.1016/j.visres.2013.12.015

Dosher, B., & Lu, Z. L. (2017). Visual perceptual learning and models. *Annual Review of Vision Science, 3*, 343-363. doi: 10.1146/annurev-vision-102016-061249

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087-1101. https://doi.org/10.1037/0022-3514.92.6.1087

Duncan, C. P., & Underwood, B. J. (1952). Retention of transfer in motor learning after 24 hours and after 14 months as a function of degree of first-task learning and inter-task similarity. *WADC Technical Report*, 99. https://doi.org/10.1016/0022-460X(71)90105-2

Edwards, J. D., Ruva, C. L., O'Brien, J. L., Haley, C. B., & Lister, J. J. (2013). An examination of mediators of the transfer of cognitive speed of processing training to everyday functional performance. *Psychology and Aging*, *28*(2), 314-321. https://doi.org/10.1037/a0030474

Fahle, M., & Morgan, M. (1996). No transfer of perceptual learning between similar stimuli in the same retinal position. *Current Biology*, *6*(3), 292–297. https://doi.org/10.1016/S0960-9822(02)00479-7

Fahle, M., & Henke-Fahle, S. (1996). Interobserver variance in perceptual performance and learning. *Investigative Ophthalmology and Visual Science*, *37*(5), 869–877.

Fiorentini, A., & Berardi, N. (1981). Learning in grating waveform discrimination: Specificity for orientation and spatial frequency. *Vision Research*, *21*(7), 1149–1158. https://doi.org/10.1016/0042-6989(81)90017-1

Fiser, J., & Lengyel, G. (2019). A common probabilistic framework for perceptual and statistical learning. *Current Opinion in Neurobiology*, *58*, 218-228. https://doi.org/10.1016/j.conb.2019.09.007

Freyer, F., Becker, R., Dinse, H. R., & Ritter, P. (2013). State-dependent perceptual learning. *The Journal of Neuroscience*, *33*(7), 2900–2907. https://doi.org/10.1523/JNEUROSCI.4039-12.2013

Fulvio, J. M., Green, C. S., & Schrater, P. R. (2014). Task-specific response strategy selection on the basis of recent training experience. *PLoS Computational Biology, 10*, e1003425. https://doi.org/10.1371/journal.pcbi.1003425

Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review*, *62*(1), 32–41. https://doi.org/10.1037/h0048826

Glass, B. D., Maddox, W. T., & Love, B. C. (2013). Real-time strategy game training: emergence of a cognitive flexibility trait. *PLoS one, 8*(8), e70350. https://doi.org/10.1371/journal.pone.0070350

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*(6939), 534–537. https://doi.org/10.1038/nature01647

Green, C.S., Pouget, A., & Bavelier, D. (2010). Improved probabilistic inference as a general learning mechanism with action video games. *Current Biology*, *20*(17), 1573-1579. https://doi.org/10.1016/j.cub.2010.07.040

Green, C. S., Kattner, F., Siegel, M. H., Kersten, D., & Schrater, P. R. (2015). Differences in perceptual learning transfer as a function of training task. *Journal of Vision, 15*(10), 5. https://doi.org/10.1167/15.10.5

Green, C. S., Banai, K., Lu, Z. L., & Bavelier, D. (2018). Perceptual learning. In J. T. Serences (Ed.), *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience* (*Vol. 2*). John Wiley & Sons, Inc.

Grossman, R., & Salas, E. (2011). The transfer of training: What really matters. *International Journal of Training and Development, 15*(2), 103–120. https://doi.org/10.1111/j.1468-2419.2011.00373.x

Harlow, H. F. (1949). The formation of learning sets. *Psychological Review, 56*(1), 51-65. https://doi.org/10.1037/h0062474

Harris, H., Gliksberg, M., & Sagi, D. (2012). Generalized perceptual learning in the absence of sensory adaptation. *Current Biology, 22*(19), 1813–1817. https://doi.org/10.1016/j.cub.2012.07.059

Heppe, H., Kohler, A., Fleddermann, M. T., & Zentgraf, K. (2016). The relationship between expertise in sports, visuospatial, and basic cognitive skills. *Frontiers in Psychology, 7*, 904. https://doi.org/10.3389/fpsyg.2016.00904

Herold, D. M., Davis, W., Fedor, D. B., & Parsons, C. K. (2002). Dispositional influences on transfer of learning in multistage training programs. *Personnel Psychology, 55*(4), 851-869. https://doi.org/10.1111/j.1744-6570.2002.tb00132.x

Herzog, M. H., & Fahle, M. (1997). The role of feedback in learning a vernier discrimination task. *Vision Research, 37*(15), 2133–2141. https://doi.org/10.1016/S0042-6989(97)00043-6

Holmes, J., & Gathercole, S. (2014). Taking working memory training from the laboratory into schools. *Educational Psychology, 34*(4), 440-450. https://doi.org/10.1080/01443410.2013.797338

Hung, S. C., & Seitz, A. R. (2014). Prolonged training at threshold promotes robust retinotopic specificity in perceptual learning. *The Journal of Neuroscience, 34*(25), 8423–8431. https://doi.org/10.1523/JNEUROSCI.0745-14.2014

Jacobs, D. M., Michaels, C. F., & Runeson, S. (2000). Learning to perceive the relative mass of colliding balls: The effects of ratio scaling and feedback. *Perception & Psychophysics, 62*(7), 1332–1340. https://doi.org/10.3758/BF03212135

Jeter, P. E., Dosher, B. A., & Lu, Z. (2009). Task precision at transfer determines specificity of perceptual learning. *Journal of Vision, 9*(1), 1–13. https://doi.org/10.1167/9.3.1

Jeter, P. E., Dosher, B. A., Liu, S.-H., & Lu, Z.-L. (2010). Specificity of perceptual learning increases with increased training. *Vision Research*, *50*(19), 1928–1940. https://doi.org/10.1016/j.visres.2010.06.016

Karbach, J., & Unger, K. (2014). Executive control training from middle childhood to adolescence. *Frontiers in Psychology, 5*, 390. https://doi.org/10.3389/fpsyg.2014.00390

Karni, A., Tanne, D., Rubenstein, B. S., Askenasy, J. J. M., & Sagi, D. (1994). Dependence on REM sleep of overnight improvement of a perceptual skill. *Science*, *265*(29), 679–682. doi: 10.1126/science.8036518

Kattner, F., Cochrane, A., & Green, C. S. (2017). Trial-dependent psychometric functions accounting for perceptual learning in 2-AFC discrimination tasks. *Journal of Vision*, *17*(11), 3. https://doi.org/10.1167/17.11.3

Kattner, F., Cochrane, A., Cox, C. R., Gorman, T. E., & Green, C. S. (2017). Perceptual learning generalization from sequential perceptual training as a change in learning rate. *Current Biology*, *27*(6), 840–846. https://doi.org/10.1016/j.cub.2017.01.046

Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2010). Learning to learn causal models. *Cognitive Science*, *34*(7), 1185–1243. https://doi.org/10.1111/j.1551-6709.2010.01128.x

Kimchi, R., & Palmer, S. E. (1982). Form and texture in hierarchically constructed patterns. *Journal of Experimental Psychology: Human Perception and Performance, 8*(4), 521-535. https://doi.org/10.1037/0096-1523.8.4.521

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3? *Perception, 36,* (ECVP Abstract Supplement).

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"?. *Psychological Science, 19*(6), 585-592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Large, A., Bediou, B., Cekic, S., Hart, Y., Bavelier, D., & Green, C.S. (2019). Cognitive and behavioral correlates of achievement in a complex multi-player video game. *Media and Communication, 7*(4), 198-212. https://doi.org/10.17645/mac.v7i4.2314

Liu, Z., & Weinshall, D. (1999). Mechanisms of generalization in perceptual learning. *Vision Research, 40*(1), 97-109. doi: 10.1016/s0042-6989(99)00140-6.

Lu, Z. L., & Dosher, B. A. (2009). Mechanisms of perceptual learning. *Learning & Perception, 1*(1), 19-36. https://doi.org/10.1556/lp.1.2009.1.3

Machin, M. A., & Fogarty, G. J. (2003). Perceptions of training-related factors and personal variables as predictors of transfer implementation intentions. *Journal of Business and Psychology, 18*(1), 51-71. https://doi.org/10.1023/A:1025082920860

Maniglia, M., & Seitz, A. R. (2018). Towards a whole brain model of perceptual learning. *Current Opinion in Behavioral Sciences, 20*, 47-55. https://doi.org/10.1016/j.cobeha.2017.10.004

Martin, A., Barnes, K. A., & Stevens, W. D. (2012). Spontaneous neural activity predicts individual differences in performance. *Proceedings of the National Academy of Sciences, 109*(9), 3201–3202. https://doi.org/10.1073/pnas.1200329109

Naquin, S. S., & Holton III, E. F. (2002). The effects of personality, affectivity, and work commitment on motivation to improve work through learning. *Human Resource Development Quarterly, 13*(4), 357-376. https://doi.org/10.1002/hrdq.1038

Ophir, E., Nass, C., & Wagner, A. D. (2009). Cognitive control in media multitaskers. *Proceedings of the National Academy of Sciences, 106*(37), 15583-15587. https://doi.org/10.1073/pnas.0903620106

Poggio, T., Fahle, M., & Edelman, S. (1991). Fast perceptual learning in visual hyperacuity. *Science, 256*(5059), 1018-1021. doi: 10.1126/science.1589770

Pugh, K. J., & Bergin, D. A. (2006). Motivational influences on transfer. *Educational Psychologist, 41*(3), 147–160. https://doi.org/10.1207/s15326985ep4103_2

Powers, K. L., Brooks, P. J., Aldrich, N. J., Palladino, M. A., & Alfieri, L. (2013). Effects of video-game play on information processing: A meta-analytic investigation. *Psychonomic Bulletin & Review, 20*(6), 1055-1079. doi:10.3758/s13423-013-0418-z

Richardson, M., & Abraham, M. (2009). Conscientiousness and achievement motivation predict performance. *European Journal of Personality, 23*, 589–605. https://doi.org/10.1002/per

Ritchie, S. J., & Tucker-Drob, E. M. (2018). How much does education improve intelligence? A meta-analysis. *Psychological Science, 29*(8), 1358-1369. https://doi.org/10.1177/0956797618774253

Rogers, R. D., & Monsell, S. (1995). Costs of a predictible switch between simple cognitive tasks. *Journal of Experimental Psychology: General, 124*(2), 207-231.

Rohde, T. E., & Thompson, L. A. (2007). Predicting academic achievement with cognitive ability. *Intelligence, 35*(1), 83-92. https://doi.org/10.1016/j.intell.2006.05.004

Ross, L. A., Edwards, J. D., O'Connor, M. L., Ball, K. K., Wadley, V. G., & Vance, D. E. (2016). The transfer of cognitive speed of processing training to older adults' driving mobility across 5 years. *Journals of Gerontology: Series B, 71*(1), 87-97. https://doi.org/10.1093/geronb/gbv022

Sagi, D. (2011). Perceptual learning in vision research. *Vision Research, 51*(13), 1552–1566. https://doi.org/10.1016/j.visres.2010.10.019

Salthouse, T. A. (1993). Speed mediation of adult age differences in cognition. Developmental Psychology, 29(4), 722-738. https://doi.org/10.1037/0012-1649.29.4.722

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103*(3), 403-428.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207–217. https://doi.org/10.1111/j.1467-9280.1992.tb00029.x

Schubert, T., Finke, K., Redel, P., Kluckow, S., Müller, H., & Strobach, T. (2015). Video game experience and its influence on visual attention parameters: An investigation using the framework of the Theory of Visual Attention (TVA). *Acta Psychologica, 157*, 200-214. https://doi.org/10.1016/j.actpsy.2015.03.005

Schultz, R., Alderton, D., & Hyneman, A. (2011). *Individual Differences and Learning Performance in Computer-based Training* (No. NPRST-TN-11-4). NAVY PERSONNEL RESEARCH STUDIES AND TECHNOLOGY MILLINGTON TN

Seitz, A. R. (2017). Perceptual learning. *Current Biology, 27*(13), R631-R636. https://doi.org/10.1016/j.cub.2017.05.053

Shibata, K., Sagi, D., & Watanabe, T. (2014). Two-stage model in perceptual learning: Toward a unified theory. *Annals of the New York Academy of Sciences, 1316,* 18-28. doi: 10.1111/nyas.12419

Snell, N., Kattner, F., Rokers, B., & Green, C. S. (2015). Orientation transfer in vernier and stereoacuity training. *PloS one, 10*(12), e0145770. https://doi.org/10.1371/journal.pone.0145770

Titz, C., & Karbach, J. (2014). Working memory and executive functions: Effects of training on academic achievement. *Psychological Research, 78*(6), 852-868. https://doi.org/10.1007/s00426-013-0537-1

Toril, P., Reales, J. M., & Ballesteros, S. (2014). Video game training enhances cognition of older adults: A meta-analytic study. *Psychology and Aging, 29*(3), 706-716. doi: 10.1037/a0037507

Torrubia, R., Avila, C., Moltó, J., & Caseras, X. (2001). The Sensitivity to Punishment and Sensitivity to Reward Questionnaire (SPSRQ) as a measure of Gray's anxiety and impulsivity dimensions. *Personality and Individual Differences, 31*(6), 837-862. https://doi.org/10.1016/S0191-8869(00)00183-5

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent?. *Journal of Memory and Language*, *28*(2), 127-154. https://doi.org/10.1016/0749-596X(89)90040-5

Uttal, D. H., Miller, D. I., & Newcombe, N. S. (2013). Exploring and enhancing spatial thinking: Links to achievement in science, technology, engineering, and mathematics?. *Current*

Directions in Psychological Science, 22(5), 367-373. https://doi.org/10.1177/0963721413484756

Ventura, M., Shute, V., & Zhao, W. (2013). The relationship between video game use and a performance-based measure of persistence. *Computers & Education*, *60*(1), 52-58. https://doi.org/10.1016/j.compedu.2012.07.003

Wang, P., Liu, H. H., Zhu, X. T., Meng, T., Li, H. J., & Zuo, X. N. (2017). Action video game training for healthy adults: A meta-analytic study. *Frontiers in Psychology, 7*, 907. doi: 10.3389/fpsyg.2016.00907

Wang, R., Zhang, J. Y., Klein, S. A., Levi, D. M., & Yu, C. (2012). Task relevancy and demand modulate double-training enabled transfer of perceptual learning. *Vision Research, 61*, 33–38. https://doi.org/10.1016/j.visres.2011.07.019

Watanabe, T., & Sasaki, Y. (2015). Perceptual learning: Toward a comprehensive theory. *Annual Review of Psychology*, *66*, 197–221. https://doi.org/10.1016/j.biotechadv.2011.08.021

Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of Personality and Social Psychology*, *54*(6), 1063-1070.

Withagen, R., & van Wermeskerken, M. (2009). Individual differences in learning to perceive length by dynamic touch: Evidence for variation in perceptual learning capacities. *Attention, Perception & Psychophysics*, *71*(1), 64–75. https://doi.org/10.3758/APP

Xiao, L. Q., Zhang, J. Y., Wang, R., Klein, S. A., Levi, D. M., & Yu, C. (2008). Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology, 18*(24), 1922–1926. https://doi.org/10.1016/j.cub.2008.10.030

Zhang, J. Y., Zhang, G. L., Xiao, L. Q., Klein, S. A, Levi, D. M., & Yu, C. (2010). Rule-based learning explains visual perceptual learning and its specificity and transfer. *Journal of Neuroscience, 30*(37), 12323–12328. https://doi.org/10.1523/JNEUROSCI.0704-10.2010

Appendix 9. Individual differences in Spatial Span and Matrix Reasoning

Model formula

clickAcc ~ inv_logit(thAsym) + (inv_logit(thStart) - inv_logit(thAsym)) * (2^((1 - trialNum)/(2 + 2^thRate))); thAsym ~ fb + orderCat + (SS || subID); thStart ~ fb + orderCat + (SS || subID); thRate ~ fb + orderCat + (1 || subID)

Fixed Effects

	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat
thAsym_Intercept	1.38	0.17	1.07	1.73	1.01
thAsym_fbn	-0.22	0.23	-0.67	0.24	1.00
thAsym_orderCat	0.02	0.23	-0.42	0.52	1.00
thStart_Intercept	1.21	0.17	0.90	1.57	1.00
thStart_fbn	-0.19	0.22	-0.62	0.23	1.00
thStart_orderCat	-0.33	0.23	-0.79	0.13	1.00
thRate_Intercept	2.53	1.54	-0.65	5.35	1.02
thRate_fbn	1.41	1.54	-1.64	4.44	1.00
thRate_orderCat	2.22	1.60	-1.03	5.30	1.00

Overall accuracy on set sizes

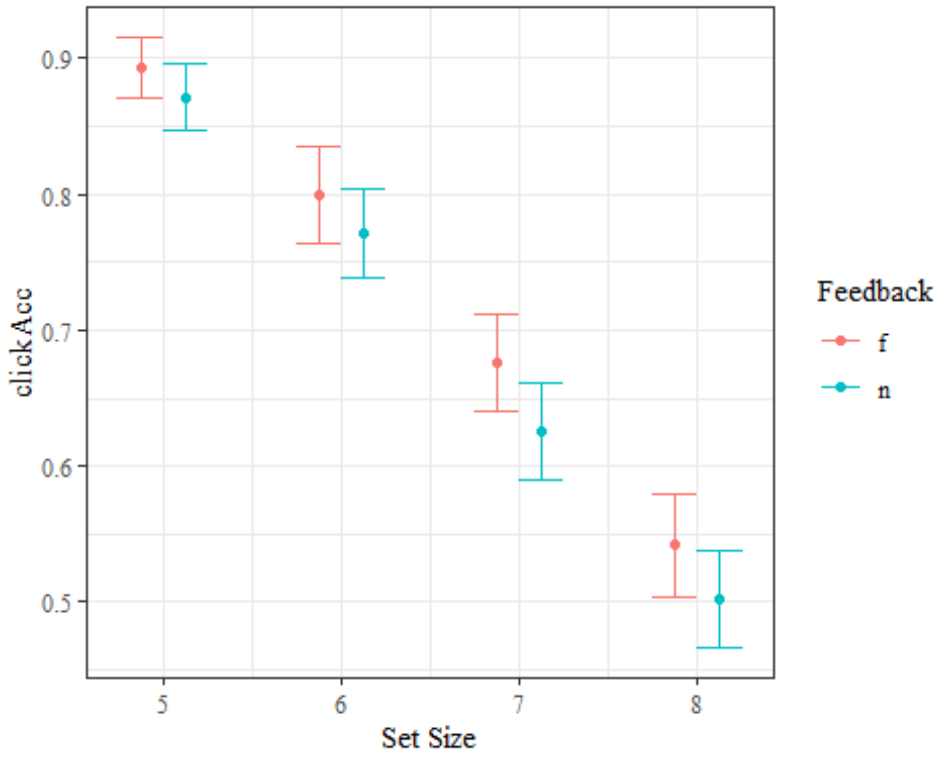


Figure A3. Feedback was systematically associated with better overall performance.

Dividing results by feedback condition

Without-feedback:

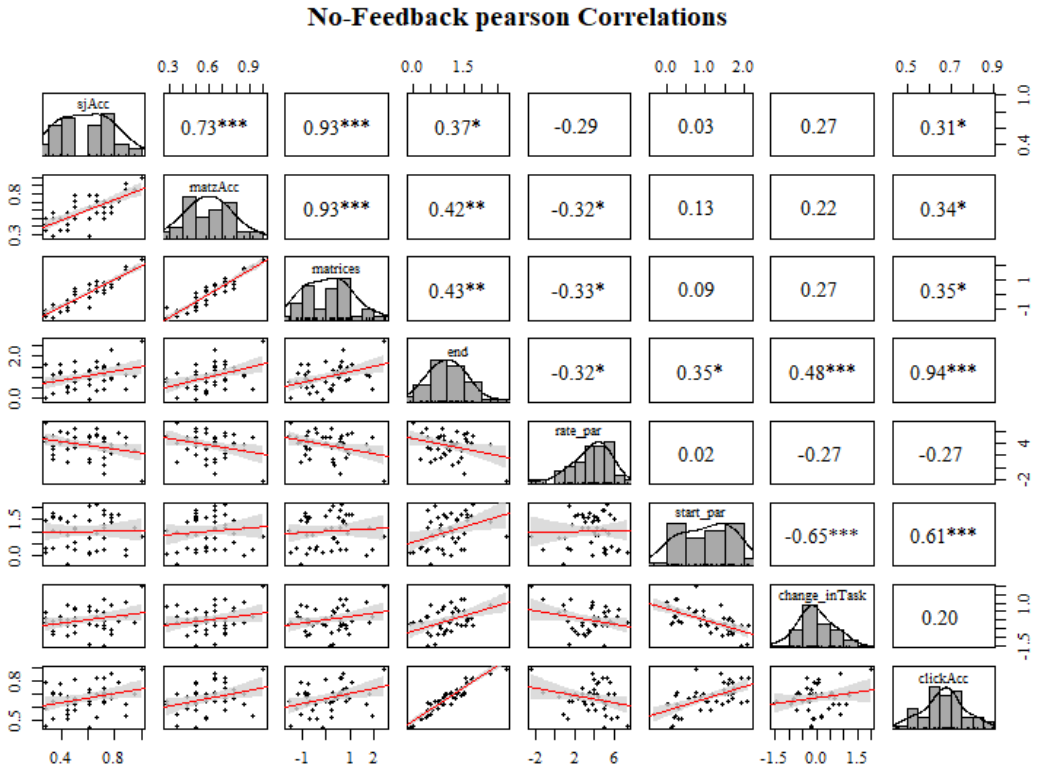


Figure A4. Product-moment correlations between variables from participants not receiving feedback in Study 5.2.

With-feedback:

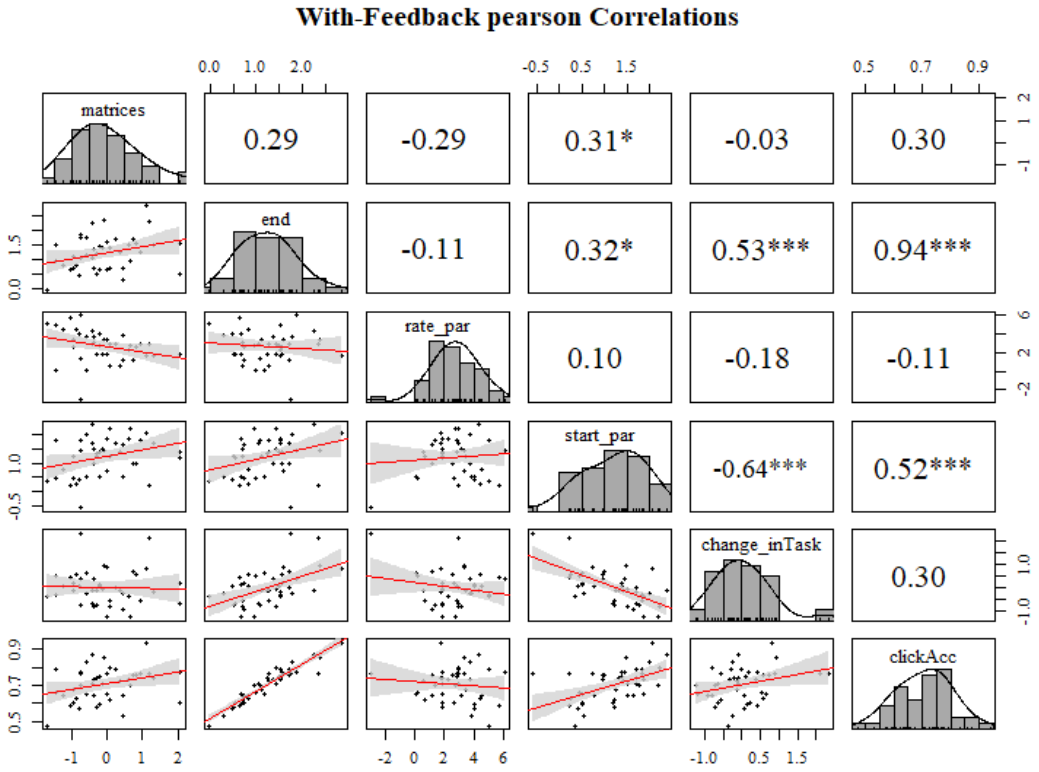


Figure A5. Product-moment correlations between variables from participants receiving feedback in Study 5.2.

Splitting Study 5.2 by people whose scores increased vs. decreased over time

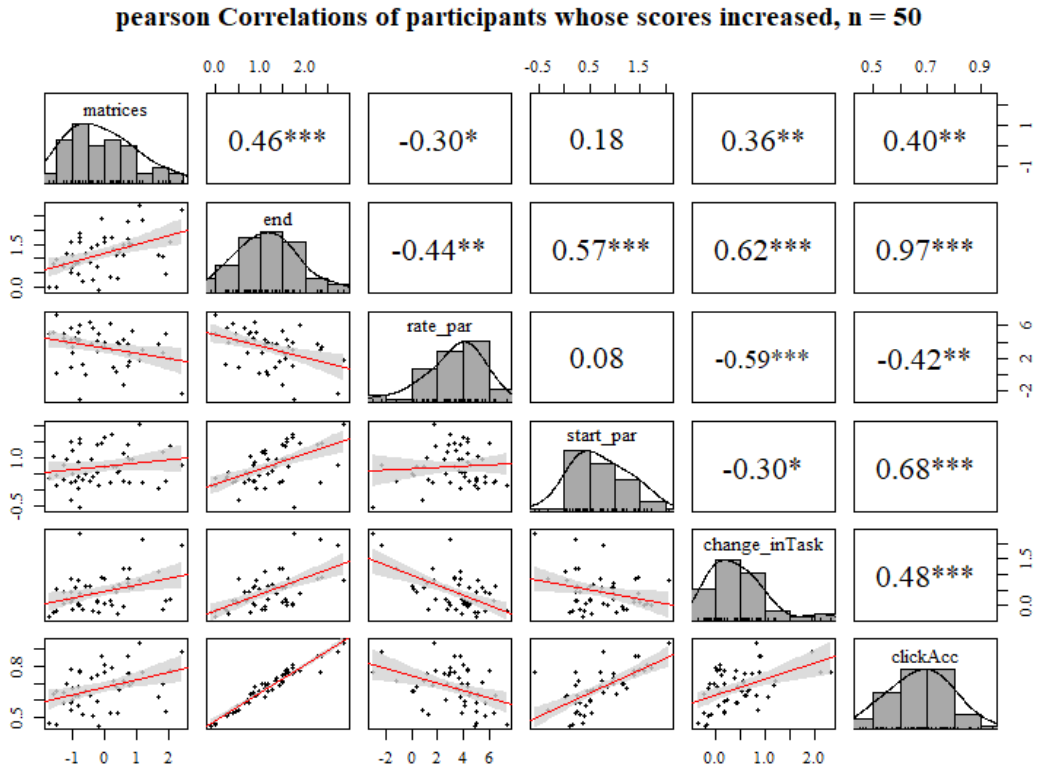


Figure A6. Of the participants whose scores increased, 27 were in the feedback-absent condition and 23 were in the feedback-present condition.

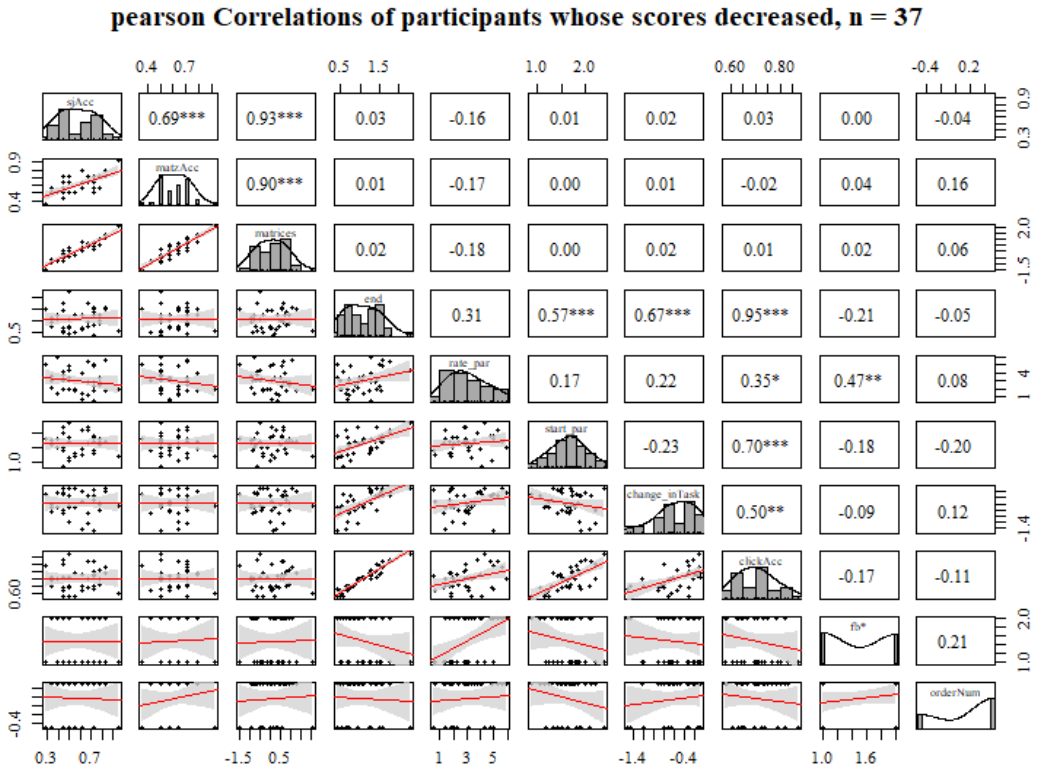


Figure A7. These participants' data appear to have none of the reliable associations observed previously, indicating that the effects are largely driven by individuals who improved on the task. Of the participants whose scores decreased, 18 were in the feedback-absent condition and 19 were in the feedback-present condition.

Appendix 10. Cochrane (2020)

The Journal of Open Source Software

249

TEfits: Nonlinear regression for time-evolving indices

Aaron Cochrane[1]

1 University of Wisconsin - Madison

DOI: 10.21105/joss.02535

Software

- Review
- Repository
- Archive

Editor: Christopher R. Madan

Reviewers:

- @ejhigson
- @paul-buerkner

Submitted: 21 May 2020
Published: 19 August 2020

Summary

Within behavioral science, it is common for data to be paradigmatically collected through repeated measurement of behavior (e.g., on each of 400 trials a human presses one of two buttons to indicate which of two possible stimuli they saw). Typical analytic tools used alongside such designs, such as ANOVA, linear regression, or T tests, implicitly assume that the data arises from distributions that are stationary across the repeated individual measurements (i.e., that every trial is independently and identically sampled from the same distribution, or *iid*, conditional on experimentally manipulated or observed variables). Interestingly, the use of such analytic tools is common even in those areas of behavioral science that are inherently concerned with time-evolving changes in behavior such as learning, memory, priming, adaptation, vigilance, cognitive control. For instance, in learning research it is common for researchers to first divide the repeated measurements into temporal bins (e.g., trials 1-100; 101-200; 201-300). They then calculate means within those bins, before applying the analysis tools above. Such an analytic method is explicitly modelling a process where performance can change between, but not within, bins. That is, conditional stationarity is assumed within the temporal bins. Beyond these fields, research methods in many others (e.g., attention, development, neuroscience, perception) attempt to by-pass the problem of non-stationarity by utilizing practice trials prior to collecting behavioral data. Practice trials are intended to give participants enough practice with the task that they reach a stable level of performance. However, whether this assumption is true is rarely tested. Our previous work has demonstrated, across several different experimental contexts, that by-trial modeling of performance provides estimates of the full timecourse of behavioral change. In doing so, these models of nonlinear monotonic *trend* stationarity provide both better estimates of behavior, as well as allowing for deeper inferences regarding the underlying processes at work, than statistical methods that assume that behavioral data remains unconditionally stationary over the course of a set of measurements (Kattner, Cochrane, & Green, 2017).

`TEfits` is a `R` package for fitting and assessing time-evolving models to data common in behavioral science. `TEfits` is designed with behavioral science researchers with a range of interests and expertise in models of time-dependent changes in behavior. Although many excellent nonlinear regression methods exist in `R`, most notably using the powerful and flexible Bayesian package `brms` (Buerkner, 2017), but also including functions such as `nls` from the core `R` `stats` package, these methods can be difficult to learn and integrate into the workflow of researchers not familiar with nonlinear regression. The user-oriented functions of `TEfits` are designed to be friendly to `R` users with minimal experience implementing nonlinear models. Extensions of this base functionality allow for simple use of various time-evolving indices (e.g., psychometric function threshold or d prime), objective functions, and/or functional forms of time-related change. Default constraints are applied to models for stability and reproducibility, but boundaries on parameters or predicted values are fully user-defineable. `TEfits` is designed to operate with minimal dependencies on other `R` packages. However, certain functions allow for optional simulation from model fits using `MASS` (Venables & Ripley, 2002), fitting of hierarchical models using `lme4` (Bates, Machler, Bolker, & Walker, 2015), or re-fitting a model using Bayesian methods using `brms` (Buerkner, 2017).

`TEfits` is being actively used in learning and memory research, with several manuscripts in preparation or under review, and results using `TEfits` have been presented at academic conferences. Primary areas of use to-date include assessments of the most appropriate learning functional form of visual perceptual improvements, testing for learning and generalization in the field of radiological diagnosis (Johnston et al., 2020), and modeling rapid shifts of attentional control in response to environmental statistics. However, the use of `TEfits` could be appropriate in any domain where individuals repeatedly engage with the same task (i.e., most psychological tasks). The ease of use and wide applicability of `TEfits` models should remove barriers from many behavioral scientists' assessment of the assumption of stationarity. Instead of this assumption users are provided a framework for understanding the changes that occur in nonstationary (i.e., *iid* conditioned on a time-evolving trend) distributions of behavioral data.

Funding and Support

This work has been supported in part by US Office of Naval Research Grant ONR-N000141712049.

References

Bates, D., Machler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). doi:10.18637/jss.v067.i01

Buerkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). doi:10.18637/jss.v080.i01

Johnston, I., Ji, M., Cochrane, A., Demko, Z., Robbins, J., Stephenson, J., & Green, C. S. (2020). Perceptual learning of appendicitis diagnosis in radiological images. *Journal of Vision*, *20*(8). doi:10.1167/jov.20.8.16

Kattner, F., Cochrane, A., & Green, C. S. (2017). Trial-dependent psychometric functions accounting for perceptual learning in 2-AFC discrimination tasks. *Journal of Vision*, *17*(11), 3. doi:10.1167/17.11.3

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (Fourth.). New York: Springer. doi:10.1007/978-0-387-21706-2

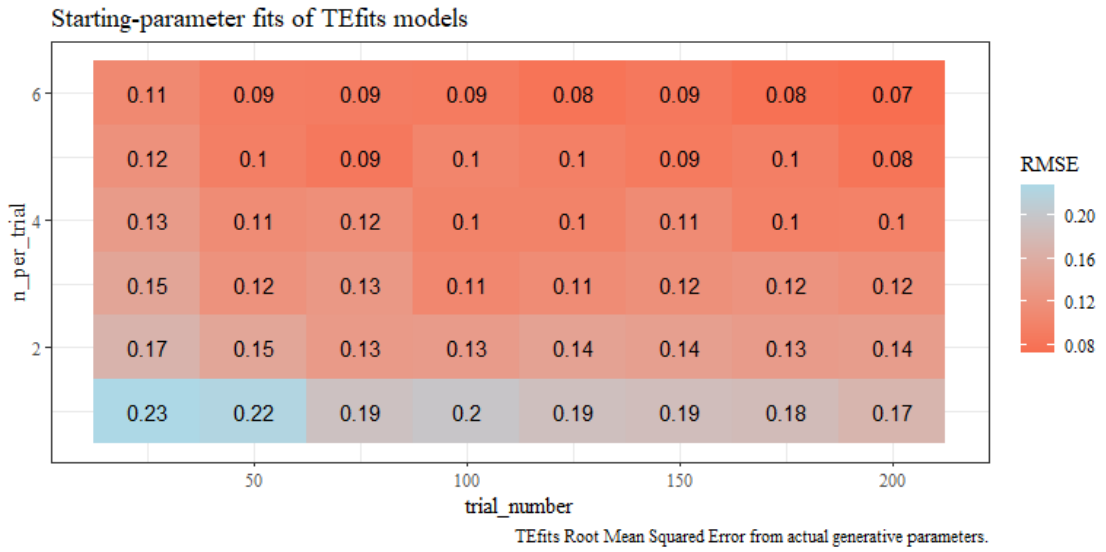
Appendix 11. TEfits performance in absolute terms

Under good conditions error is <10% of possible variation and correlation with actual parameters is around .9. But under compromised situations, parameter estimates can degrade pretty quickly.

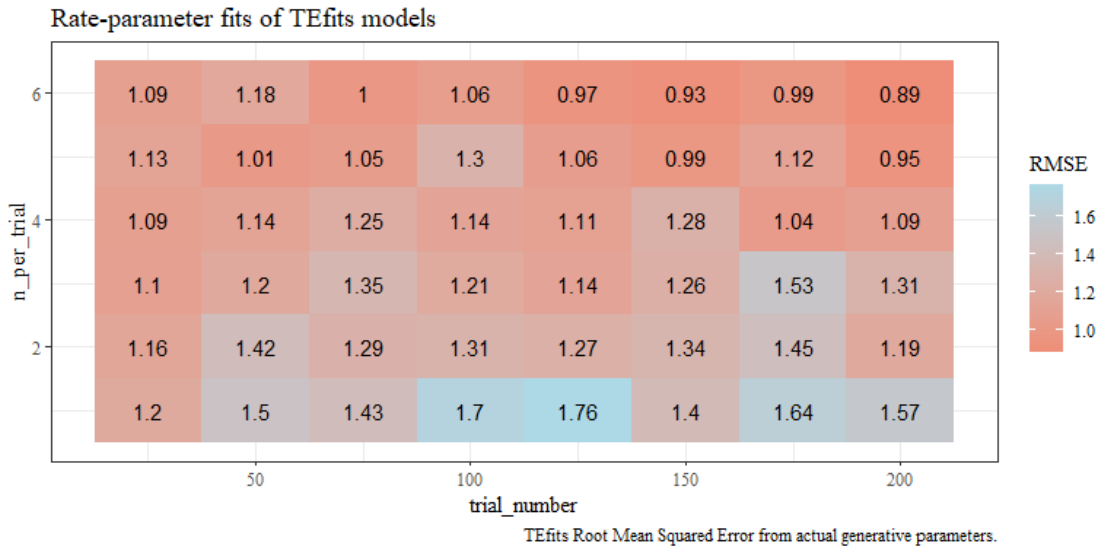
RMSE

Raw RMSE between **TEfits** parameters and true parameters.

A.



B.



C.

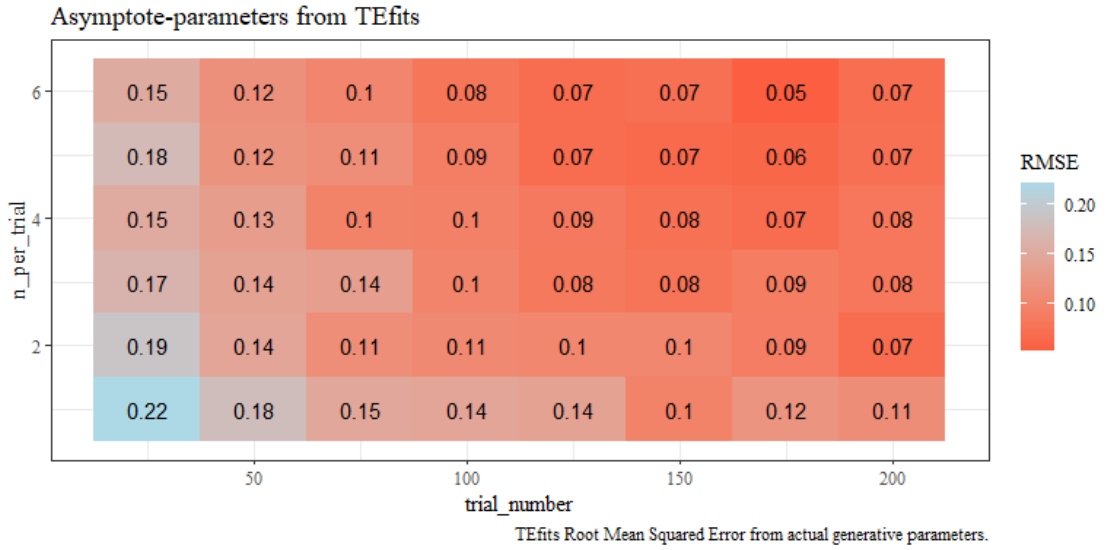
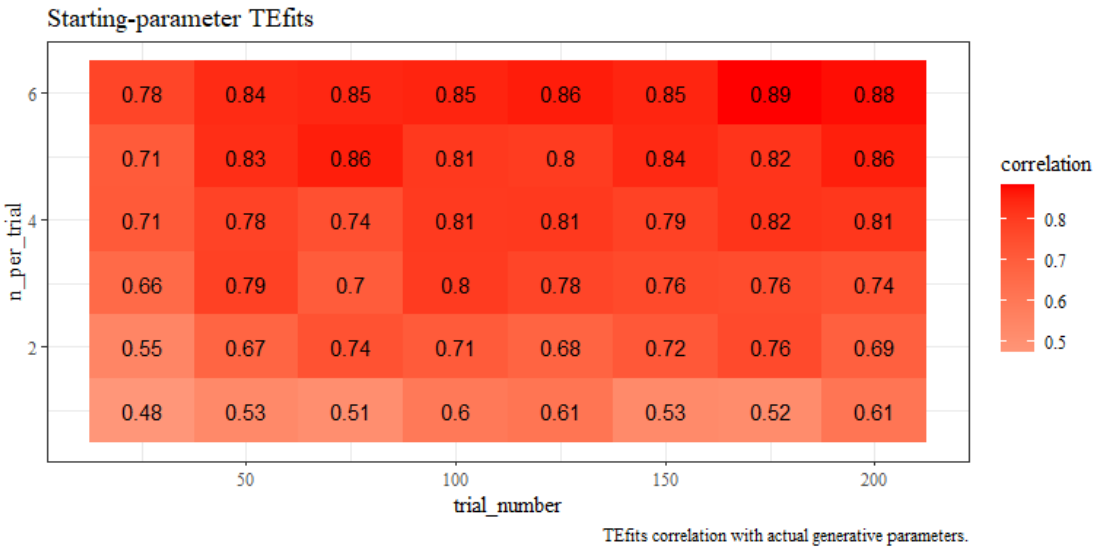


Figure A8. TEfits RMSE.

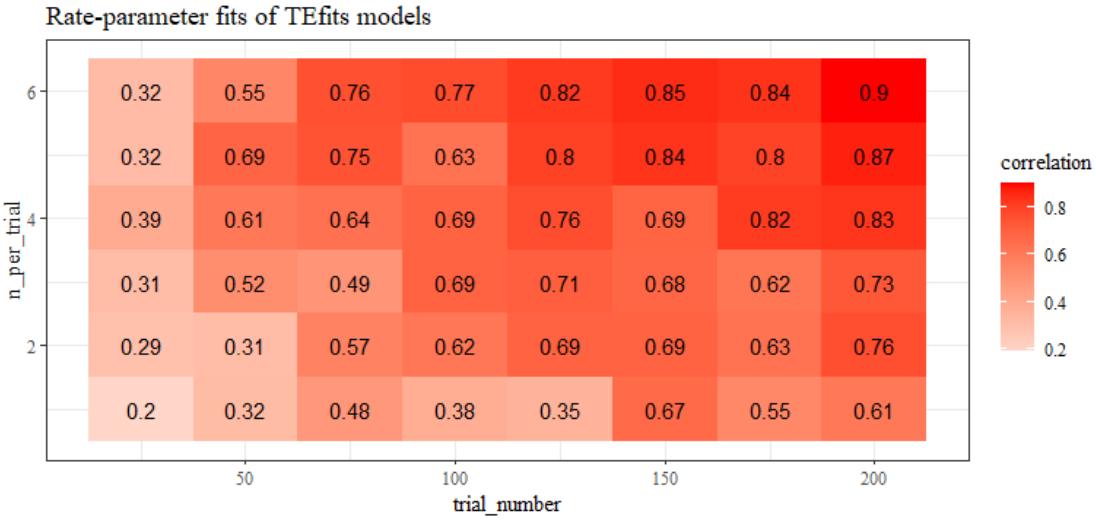
Correlations

A.



TEfits correlation with actual generative parameters.

B.



TEfits correlation with actual generative parameters.

C.

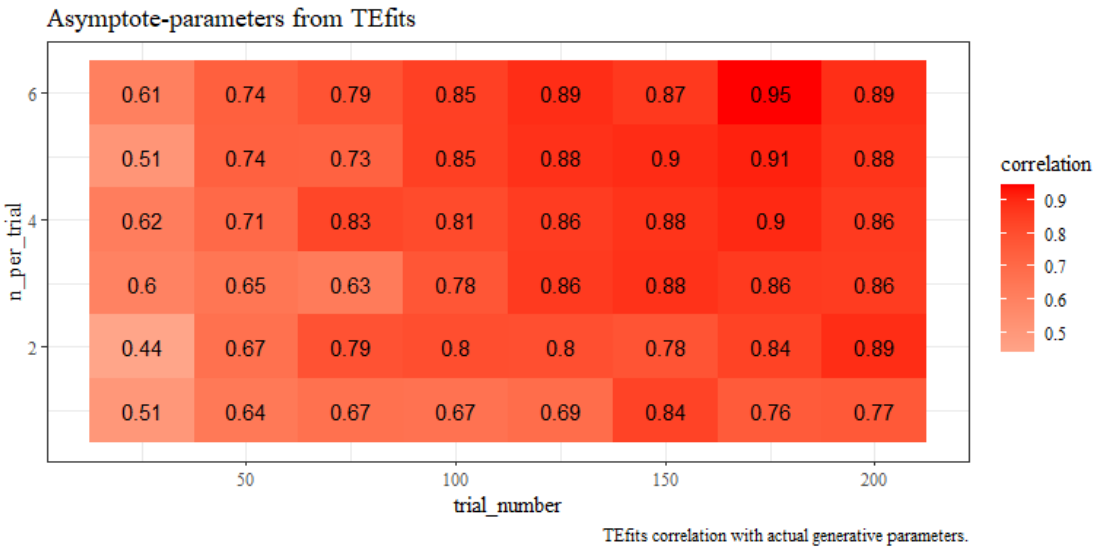


*Figure A9. Raw correlations between **TEfits** parameters and true parameters.*