Title
Technical Challenges in Perceptual Learning Research

Author
Jacques, Theodore

Publication Date
2019

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Technical Challenges in Perceptual Learning Research

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Psychology

by

Theodore Jacques

September 2019

Dissertation Committee:
Dr. Aaron R. Seitz, Chairperson
Dr. George John Andersen
Dr. Lawrence D. Rosenblum

The Dissertation of Theodore Jacques is approved:

Committee Chairperson

University of California, Riverside

Dedication and Acknowledgements

I am grateful for the support and assistance of everyone who has contributed to my professional development over the years. Additionally, I give a special thanks to my mother Dr. Jane Jacques and my father Thomas Jacques. Finally, on a personal note, as a professional I have always acted as though my work is independent of my personal life and feelings, science is about understanding the world and not about satisfying a personal agenda. Nevertheless, in this instance I will take the opportunity to express my faith and so this work is dedicated to God.

ABSTRACT OF THE DISSERTATION

Technical Challenges in Perceptual Learning Research

by

Theodore Jacques

Doctor of Philosophy, Graduate Program in Psychology
University of California, Riverside, September 2019
Dr. Aaron R. Seitz, Chairperson

There is considerable variability in the observed effects of perceptual learning, and consequently considerable debate about the underlying mechanisms of learning. Resolving these irregularities and debates requires careful experimental control and precise application of research methods. At times, improving consistency between labs and methods is the best solution, but sometimes the best choice is to develop a new approach. Here we describe first a traditional perceptual learning experiment aimed at understanding individual differences. Then, when it was discovered that important data-analytic assumptions were violated by that experiment, we describe the development of novel data analytic techniques to help understand the results. Finally, we discuss the development of an entirely new paradigm to help understand aspects of perceptual learning that are at present difficult to measure reliably.

Table of Contents

List of Figures

List of Tables

General Introduction

A good research question can be articulated clearly and simply in a way that even non-experts can understand. What factors influence a person's ability to learn a task? This simple phrasing outlines a focus on the underlying mechanisms of learning, and communicates an intellectual perspective that views learning as a process whose rules can be understood. In practice as we unpack this question experimentally we recognize it as a multi-faceted problem encompassing a variety of aspects of the learning experience such as pre-existing aptitude in the relevant cognitive domains (qualities of the learner), the design of the learning experience itself (qualities of the task), and the ways that subjects interface with the task (strategy and task-learning). By understanding the ways that these factors interact with one another we can understand, predict, and improve the learning process for a variety of applications.

Perceptual learning provides a convenient perspective for understanding the learning process. By focusing on low-level perceptual tasks and by eliminating many of the confounds in language and knowledge that make other types of learning more complex we can concentrate more effectively on the mechanisms underlying learning. And although perceptual learning provides a useful model system for learning in general, there are a variety of direct applications for perceptual learning including treatment of patients with visual impairments (Polat et al., 2009), the training of radiologists (Sowden et al., 2000; Kellman, 2013), machine learning (Bredeche et al., 2006), and even sports (Deveau et al., 2014).

In spite of the basic sensory foundation of the stimuli used in our experiments, perceptual learning researchers have observed a wide variety of mechanisms and learning outcomes, enough to account for many aspects of the learning process and to fuel a robust theoretical discourse. Indeed, we observe significant variability in the reliability of perceptual learning (Hung & Seitz, 2014; Huang et al., 2017; Liang et al., 2015; Zhang & Yu, 2016; Xiao et al., 2008), and it is the diverse combination of the subject-specific and task-specific factors that explain such a wide variety of outcomes. Therefore, a deeper understanding of the interactions between these factors is a necessary step towards expanding our understanding of this field.

The qualities of the learner refer to individual differences in pre-existing cognitive abilities. Unfortunately, these factors are largely unaccounted-for in perceptual learning literature (Dobres & Seitz, 2010; Herzog & Fahle, 1997; Hung & Seitz, 2011). Two possible causes of this issue are the difficulty in running participants across multiple sessions and the small sample sizes that typify experiments in the field (Hung & Seitz, 2014). The qualities of the task refer to the particular methodological paradigm chosen to measure learning and the properties of the stimulus. Details of stimulus presentation vary considerably (Karni & Sagi, 1991; Harris, Gliksberg & Sagi, 2012; Wang, Cong & Yu, 2013), and systematically manipulating these details is not a frequent topic of inquiry (see Yotsumoto et al., 2009 and Deveau, 2014 for two examples), although there is recent interest in this issue in the domain of brain game development (Deveau et al., 2015, Mohammed et al., 2017). The interaction between the subject and the task is more difficult to define, but certainly any approach to addressing this question would require the ability to measure subject behaviors at very high temporal resolution such as trial-to-trial measures of learning. To date, we know of no such method in the existing literature.

This work describes my efforts to further our understanding of these important topics. It also addresses some of the ways that existing research methodologies were insufficient for addressing my research questions. Consistent research methodologies are important in understanding perceptual learning, as mentioned earlier there is large variability in the observed reliability of perceptual learning and one example is the finding of task-specific disruption of perceptual learning (Seitz et al., 2005). Following an unsuccessful replication attempt by a different group (Aberg & Herzog, 2010), the original lab subsequently replicated both their own original results and the failure to replicate (Hung & Seitz, 2011). It is likely that the differences in outcome between the first two experiments were due to subtle methodological changes in the administration of the task between labs. This example illustrates how researchers in the same field need to be mindful of even small aspects of their task administration even when using the same task. The high variability of findings in perceptual learning research makes this a particularly important concern, especially in light of the so-called “replication crisis” in psychology (Maxwell et al., 2015). In spite of this need, it is not always possible or desirable to perfectly replicate the methods of previous researchers, in particular when the same tasks are being used to address different research questions between experiments. In such circumstances, existing methods can be inadequate and novel approaches must be developed. Consequently, this work can be thought of as a progressive departure from traditional methods as the need arose in pursuit of my research goals.

Dissertation Structure

This work is divided into three chapters. The first chapter describes a perceptual learning study conducted to understand moderating variables in perceptual learning, with the additional feature of an unusually large sample size. This chapter is representative of a typical study in the field and most directly addresses research questions in perceptual learning. Over the course of three experiments we explored the ways that differences in the task structure (roving the order of trial difficulty and the inclusion of a second training session intended to induce interference in learning) and the personal qualities of the subjects (pre-existing visual attention skill and experience with action video games) influenced learning and transfer in the well-studied Texture Discrimination Task (TDT, see Figure I.1). One of our primary findings was that roving the order of trial difficulty reduced overall performance, and this finding led to our first major problem with the standard practices in the field. This practice relates

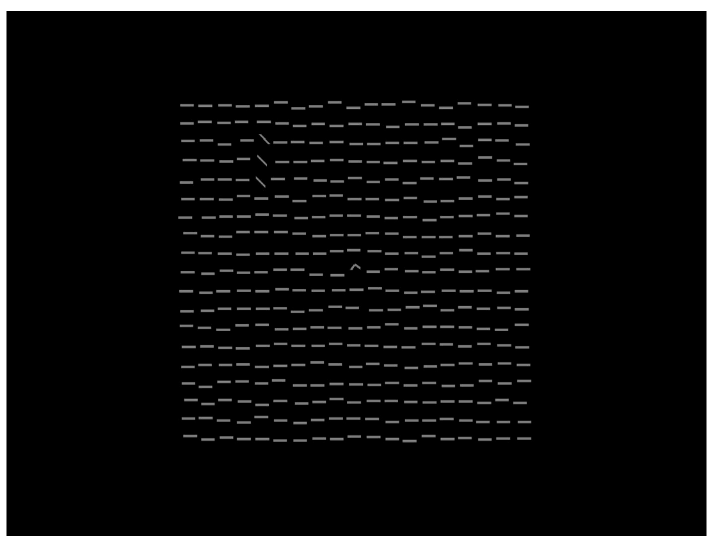


Figure I.1. Sample Texture Discrimination Task stimulus. The peripheral stimulus is located in the upper-left hand quadrant. The central target is a rotated L.

to a phenomenon known as “stimulus-independent *lapses*” (Wichmann & Hill, 2001), moments when subjects are inattentive to the task that reduce overall performance. These lapses have been known for decades, and some research has gone into understanding their impact on perceptual data (Harvey, 1986; Swanson & Birch, 1992). One of the standard methods for data analysis in the TDT is to fit the data with a psychometric function and report thresholds, and although it is possible to account for lapse rate in the psychometric function (see Prins & Kingdom, 2018 for a discussion of the topic), one of the most common ways to deal with these lapses is to simply drop subjects with low overall performance whose data could interfere with our interpretation of the results of the study.

However, we understood that the trial difficulty manipulation was partially responsible for low overall performance in our subjects, and so in this case the standard practice could not be used (the typical approach assumes that these lapses are the only reason for low baseline performance). Therefore, the second chapter deals with our departure from the traditional approach by describing this methodological concern in full detail and outlines the novel alternative method we developed for identifying these subjects. We use this approach to separate our subjects into psychometrically-fit and non psychometrically-fit groups, present additional statistical analyses of the data from the initial TDT experiment, and finally discuss the impact of these additional analyses on the conclusions drawn in chapter one.

The final chapter describes the development of an entirely new paradigm for perceptual learning research. We created this technique in response to perceived inadequacies in existing methods of estimating perceptual learning, particularly the ways that discrete testing sessions can interfere with learning and the measurement of learning over short time scales. Our novel method combines the wealth of information eye tracking can provide about human gaze behaviors with well-established image processing techniques to make inferences about subjects' learning process throughout the training procedure. Subjects are trained to locate a Gabor patch in a visual noise field, and are allowed to freely move their eyes in search of the target. The eye tracker allows us to passively observe their behavior and identify changes in perceptual bias during the course of the training. Chapter three describes the method in sufficient detail for other researchers to apply the technique to their own research questions. It also includes empirical results from a small validation study using the method that illustrates the types of data relevant to perceptual learning that eye tracking can provide.

In the service of the over-arching goal of understanding perceptual learning, it is at times necessary to blaze new trail and develop entirely new approaches. This can be necessary due to the assumptions we make about the behavior of subjects, or due to limitations in existing approaches. Our aim with this work is to demonstrate the validity of these new methods, and to improve our understanding of this important topic in cognition.

References

Aberg, K. C., & Herzog, M. H. (2010). Does perceptual learning suffer from retrograde interference? *PloS One*, *5*(12), e14161.

Bredeche, N., Shi, Z., & Zucker, J. D. (2006). Perceptual learning and abstraction in machine learning: an application to autonomous robotics. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, *36*(2), 172-181.

Deveau, J., Ozer, D. J., & Seitz, A. R. (2014). Improved vision and on-field performance in baseball through perceptual learning. *Current Biology*, *24*(4), R146-R147.

Deveau, J., Jaeggi, S. M., Zordan, V., Phung, C., & Seitz, A. R. (2015). How to build better memory training games. *Frontiers in Systems Neuroscience*, *8*, 243.

Dobres, J., & Seitz, A. R. (2010). Perceptual learning of oriented gratings as revealed by classification images. *Journal of Vision*, *10*(13), 8-8.

Harris, H., Gliksberg, M., & Sagi, D. (2012). Generalized perceptual learning in the absence of sensory adaptation. *Current Biology*, *22*(19), 1813-1817.

Harvey, L. O. (1986). Efficient estimation of sensory thresholds. *Behavior Research Methods, Instruments, & Computers*, *18*(6), 623-632.

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

Hung, S.-C., & Seitz, A. R. (2011). Retrograde interference in perceptual learning of a peripheral hyperacuity task. *PloS One*, *6*(9), e24556.

Hung, S. C., & Seitz, A. R. (2014). Prolonged training at threshold promotes robust retinotopic specificity in perceptual learning. *Journal of Neuroscience*, *34*(25), 8423-8431.

Karni, A, & Sagi, D. (1991). Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(11), 4966–70.

Kellman, P. J. (2013). Adaptive and perceptual learning technologies in medical education and training. *Military medicine*, *178*(October Supplement), 98-106.

Maxwell, S. E., Lau, M. Y. & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does “failure to replicate” really mean? *American Psychologist*, *70*(6). 487-498.

Mohammed, S., Flores, L., Deveau, J., Hoffing, R. C., Phung, C., Parlett, C. M., Sheehan, E., Lee, D., Au, J., Buschkuehl, M., Zordan, V., Jaegi, S. M. & Seitz, A. R. (2017). The benefits and challenges of implementing motivational features to boost cognitive training outcome. *Journal of Cognitive Enhancement*, *1*(4), 491-507.

Polat, U., Ma-Naim, T., & Spierer, A. (2009). Treatment of children with amblyopia by perceptual learning. *Vision research*, *49*(21), 2599-2603.

Prins, N., & Kingdom, F. A. (2018). Applying the model-comparison approach to test specific research hypotheses in psychophysical research using the Palamedes Toolbox. *Frontiers in Psychology*, *9*.

Seitz, A. R., Yamagishi, N., Werner, B., Goda, N., Kawato, M., & Watanabe, T. (2005). Task-specific disruption of perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(41), 14895–900.

Sowden, P. T., Davies, I. R., & Roling, P. (2000). Perceptual learning of the detection of features in X-ray images: a functional role for improvements in adults' visual sensitivity?. *Journal of Experimental Psychology: Human perception and performance*, *26*(1), 379.

Swanson, W. H., & Birch, E. E. (1992). Extracting thresholds from noisy psychophysical data. *Perception & Psychophysics*, *51*(5), 409-422.

Wichmann, F. A. & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293-1313.

Wang, R., Cong, L. J., & Yu, C. (2013). The classical TDT perceptual learning is mostly temporal learning. *Journal of Vision*, *13*(5), 9-9.

Yotsumoto, Y., Chang, L. H., Watanabe, T., & Sasaki, Y. (2009). Interference and feature specificity in visual perceptual learning. *Vision Research*, *49*(21), 2611–2623.

Moderating Effects of Visual Attention and

Action Video Game Play on Perceptual Learning

Theodore Jacques[1] & Aaron R. Seitz[1]

Author Affiliations:

[1]Department of Psychology, University of California, Riverside, Riverside, CA 92521

Correspondence should be addressed to:

Theodore Jacques

theodore.jacques@gmail.com

(567)-203-8232

Abstract

There is currently substantial controversy regarding the reliability of observed patterns of perceptual learning. Contributing to this controversy are a lack of accounting for individual differences and how variations in training can give rise to different patterns of learning. Here, in three experiments we investigate the impact of individual differences in attention and experience with action video game use, as well as task-specific factors, on perceptual learning in a large sample of subjects using a Texture Discrimination Task (TDT). We find a significant impact of visual attention skill but not video game use on learning and transfer in the TDT, as well as a substantial impact of the order of trial difficulty on overall performance but not learning. Of note we fail to find evidence for interference between training with different backgrounds that have been observed in previous studies with the TDT. Together these results suggest that, while differences between individuals and differences in task structure play a role in learning, previous findings on the impact of action video game use and attention in perceptual learning may be idiosyncratic to particular training circumstances.

Introduction

In recent years, there has been substantial controversy regarding the reliability of observed patterns of perceptual learning (Hung & Seitz, 2014; Huang et al., 2017; Liang et al., 2015; Zhang & Yu, 2016; Xiao et al., 2008) and in turn the mechanisms that give rise to perceptual learning (Harris, Gliksberg & Sagi, 2012; Law & Gold, 2009; Gu et al., 2011; Xiao et al., 2008, Yu, Klein & Levi, 2004). Factors contributing to variation of observed findings across studies include a lack of accounting for individual differences in learning (Dobres & Seitz, 2010; Herzog & Fahle, 1997; Hung & Seitz, 2011), the small sample sizes that typify most perceptual learning studies, and that variations in training can give rise to different patterns of learning (Hung & Seitz, 2014). Here, we address some of these issues in the context of the well-studied Texture Discrimination Task (TDT; Karni & Sagi, 1991) by examining in a large set of participants how variations in training and individual differences in attention and experience with action video game use may mediate and moderate perceptual learning.

The TDT is perhaps the best studied task in the field of perceptual learning and has provided fundamental insights regarding the degree to which learning can be specific to trained stimulus features (Sagi, 2011), the role of adaption in learning (Harris, Gliksberg & Sagi, 2012), the longevity of learning (Karni & Sagi, 1993), the roles of sleep in consolidation of learning (Censor, Karni & Sagi, 2006; Mednick, Nakayama, & Stickgold, 2003; McDevitt, Duggan, & Mednick, 2015), and the temporal phases of how learning is consolidated in the brain (Yotsumoto et al., 2009), among other findings. Notably, across studies, details of stimulus presentation during training have varied considerably (Karni & Sagi, 1991; Harris, Gliksberg & Sagi, 2012; Wang, Cong & Yu, 2013) indicating that observed thresholds and patterns of specificity depend substantially upon the details of the training regime. For example, Harris and Sagi (2012) found that adaptation can explain a substantial degree of location specificity of learning. However, their results also show much faster transfer of learning even in the adapted condition than in Karni and Sagi (1991), perhaps due to changing from a methods of limits to the method of constant stimuli in their training regime, as examined more directly by Harris and Sagi (2015). A related issue is how different stimulus features are interleaved. Yotsumoto et al. (2009) investigated how sequentially training on different target or background orientations facilitated or interfered with learning. They found, similar to prior results of training on hyperacuity stimuli (Seitz et al., 2005) or motor learning (Brashers-Krug et al., 1996), that this led to a disruption of learning. These results, combined with results from other perceptual learning paradigms (Hung & Seitz, 2011) suggest that details of how different stimulus types are distributed during training can have a significant impact on what is learned.

Other research suggests that non-stimulus specific cognitive factors can play an important role in perceptual learning. For example, in the context of the TDT, Wang et al., (2013) suggest learning can be accounted for by a narrowing of the window of temporal attention. They found that by separately training individuals to detect the orientation of masked gratings or Cs accounted for most of the learning observed from directly training on the TDT itself. Although, other data suggests that this temporal learning can be interpreted as a low-level change in temporal integration between targets and masks, rather than simply an effect of attention (Censor et al., 2009). However, the role of attention in perceptual learning is consistent with recent EEG studies showing that substantial differences in baseline alpha and alpha desynchronization occur through perceptual learning on a task that required attention to a very brief stimulus presentation period (Bays et al., 2014). Relatedly, there is a robust literature suggesting that experience with action video games both leads to improved attention abilities (Green & Bavelier, 2003; Green & Bavelier, 2012) and that action video game use influences learning on the TDT (Berard et al., 2015). Together these data suggest that differences in attention abilities may explain individual differences in both initial thresholds and learning rates.

The present study was devised to directly address these moderating and mediating factors on perceptual learning in the context of the TDT task. We assessed individual differences using a questionnaire of action video game use (Bavelier et al., 2011) and through the measurement of visual attention via the Useful Field of View (UFOV; Ball et al., 1988) task. We examined how these moderating factors interacted with factors thought to mediate learning, specifically differences in the distribution of stimulus difficulties during training and how training with different stimulus types can interfere with learning (Seitz et al., 2005; Yotsumoto et al., 2009; Berard, et al., 2015, McDevitt et al., 2015). Our sample of more than 150 participants allowed for one of the largest datasets reported on the TDT task to provide for robust results of the most consistent patterns of learning on this task.

Methods

Subjects – Undergraduate college students, 84 in Experiment 1, 51 in Experiment 2, and 85 in Experiment 3 from the University of California, Riverside (UCR) gave written consent in accord with policies of UCR's Human Subject Review Board and participated in exchange for research course credit. Of these, 10 in Experiment 1 and 13 in Experiment 2 were excluded from the analysis due to missing or incomplete data. In Experiment 3, only 42 subjects successfully completed the practice session. See Table 1.1 for details of subject age, gender, and gaming experience for those subjects who were included in the analysis. A more detailed breakdown of subject totals by experiment and condition is available in the supplemental material. Of note, performance levels varied considerably across participants and as such we also divide participants into Psychometrically-Fit (PF) and Non-Psychometrically-Fit (NPF) groups (see supplement for details, including breakdown of PF and NPF groups related to each analysis). For the main body of the paper, all subjects are included in all statistics, and PF and NPF group analyses appear only in the supplement, with the exception of mention of notable differences between conditions. Of note, low numbers of female AVGP and low numbers

	Total Subjects (number female)	Age Range	Mean Age (standard deviation)	NVGP (number female)	"in between" (number female)	AVGP (number female)
Experiment 1	74 (32)*	18-26	19.51 (1.60)	24 (16)	23 (10)	27 (6)
Experiment 2	38 (13)	18-24	19.74 (1.52)	9 (7)	11 (4)	18 (2)
Experiment 3	42 (14)	18-24	19.36 (1.57)	8 (4)	15 (8)	19 (2)

* One participant declined to indicate both gender and age

Table 1.1. Summary breakdown for participant demographics for each experiment.

of male NVGP are commonly found in studies of action video game players, which also may impact some analyses. Subjects were screened to ensure that they had not participated in any previous experiment involving a texture discrimination task. All subjects had normal or corrected-to-normal vision and reported being able to see the stimuli with no difficulty. Subjects' video game playing habits were not used as selection criterion for the study. Experiment 3 did include an exclusion criterion not present in Experiments 1 and 2. In Experiments 1 and 2 subjects were allowed unlimited practice sessions (time allowed), while in Experiment 3 subjects who could not reach adequate levels of performance in the TDT practice sessions within 5 attempts were excluded from the study (notably 43 subjects, about half of the recruited sample, were unable to pass this practice that was targeted to ensure that participants could perform the task well enough to increase the likelihood of a valid 70% threshold similar to the subjects reported in Yotsumoto et al., 2009).

Apparatus - An Apple Mac Mini running Matlab (Mathworks, Natick, MA) and Psychtoolbox Version 3 (Brainard, 1997; Pelli, 1997) was used for stimulus generation and experiment control. In Experiment 1, stimuli were displayed on a 16-inch Viewsonic PF817 monitor at a resolution of 1400x1050 pixels at 100 Hz by an NVIDIA GeForce 9400 graphics card (NVIDIA Corporation, Santa Clara, CA). In Experiments 2 and 3, stimuli were displayed on a 16-inch NEC FP2141SB monitor at a resolution of 1600x1200 pixels at 100 Hz by an NVIDIA GeForce 9400 graphics card.

Procedure - All three experiments followed the same procedure. On day 1, subjects began by completing the video game questionnaire and the UFOV task before a brief practice session on the TDT. Then subjects completed either one or two sessions of TDT task, depending on their experimental condition. On day 2, all subjects conducted two sessions of the TDT task, in Experiments 1 and 2 subjects got an additional practice session with the TDT on the second day, while subjects in Experiment 3 did not complete the practice on day 2.

UFOV Task - To measure visual attention we used a Useful Field of View (UFOV) task similar to that described by Ball et al. (1988). This display (see Figure 1.1) consisted of a target array followed by a mask stimulus (presented for 320ms). In each trial, a fixation point was displayed and subjects initiated each trial with a keypress. The target array contained a central stimulus (smiling cartoon face, long or short hair) and a set of hollow white squares spaced at 45° intervals around the screen in three rings of eight items each presented at 6.67°, 13.33°, and 20° eccentricity. The display size was 20°, viewed at a distance of 40cm (Experiment 1) or 39cm (Experiments 2,3). The target stimulus consisted of a white star located in either the 6.67° or the 20° ring. Subjects responded using the keyboard to indicate the length of the hair of the central stimulus and the mouse to indicate the location of the peripheral stimulus.

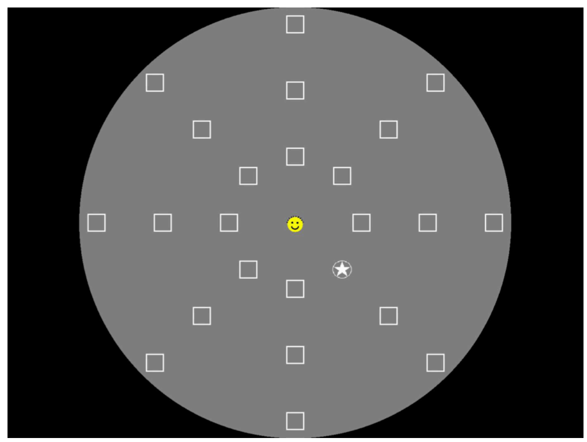


Figure 1.1. Useful Field of View task stimulus. The white star (target) is located in the inner ring and replaces the distractor box at that location.

The duration of the target array varied according to 4 independent randomly interleaved 4-down, 2-up staircases (2 starting at an initial value of 300ms; 2 starting at 100ms; one set for each target eccentricity) with consecutive correct trials (e.g. streaks) allowing for sequential steps in cases of consistent performance. For the first 16 trials only the 300ms staircases were active. The task terminated when all 4 staircases had reached a total of 8 reversals, or if each staircase had presented 72 trials. Prior to beginning the task, subjects were shown an example stimulus array, and then given two practices. The first consisted of the central target only; the second included the peripheral target with no distractors (white squares).

TDT Task - We modeled the TDT task as closely as possible to that described by Yotsumoto et al. (2009). In Experiment 1 (our Roving condition), stimulus difficulties were randomly intermixed throughout training; in Experiments 2 and 3 (our Sequential conditions) different stimulus difficulties were presented in sequential blocks of increasing difficulty. In each Experiment, subjects were randomly assigned to one of two conditions. In the interference condition, on the first day participants trained first using either vertical or horizontal background bars (A), then after a brief break trained a second time using the orthogonal background stimulus (B). Because we counterbalanced the initially-trained orientation across participants, we use the labels A and B to reflect "first trained" and "second trained" rather than "vertical" or "horizontal" background lines. Both sessions were repeated on the second day, thus we refer to the interference condition as the AB-AB group. The non-interference group trained only one session (A) on day 1, and then on the second day trained first on A and then on the orthogonal background (B). Therefore we call the non-interference group the A-AB group. When referring to specific sessions for a given participant, we will use the terminology “Day 1-A” or “A_1” to refer to the first session on the first day and so on.

In each trial, subjects first viewed a fixation point for 1000ms and were then presented with a 19x19 target array (20ms duration) followed by a mask (100ms duration) and were asked to judge the orientation (vertical or horizontal) of a peripheral target (see Figure 1.2) composed of three bars oriented 45° offset from a background of vertical or horizontal bars (0.73°x0.13°, randomly jittered by up to 0.2° around an array of points spaced by 1°). The target array was centered on the element closest to a point randomly chosen within 5-9° from the center of the display in either the upper left or upper left quadrant (randomly assigned per subject). The display size was 19°x19° of visual angle, viewed at a distance of 60cm (Experiment 1) or 69cm (Experiments 2,3). Subjects also completed a central task discriminating between a “T” and an “L” presented in place of the central element of the target array. The time interval between the target array and mask (SOA; 180, 160, 140, 120, 100, 80, and 60ms) varied either randomly (7 SOAs, each presented

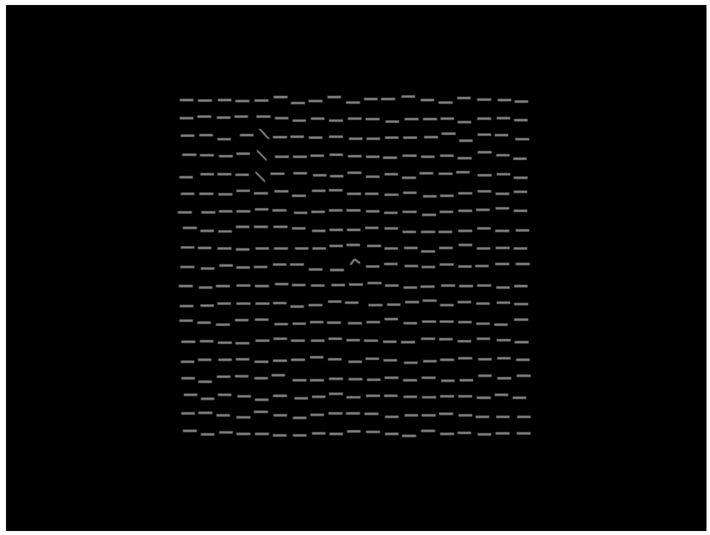


Figure 1.2. Texture Discrimination Task stimulus. The vertically-oriented (peripheral target) stimulus is located in the upper-left hand quadrant. The central target is a rotated L.

once every 7 trials in a random order, Experiment 1) or increased incrementally by block (Experiments 2,3). The mask patterns were 19 x 19 arrays of randomly oriented "v" shapes. Subjects were given up to 2000ms to respond via keyboard, and received auditory feedback on the central task only. All sessions consisted of 273 trials divided into 39 mini-blocks of 7 trials of random SOAs (Experiment 1) or 7 blocks of 39 trials of increasing difficulty (Experiments 2,3). Prior to training on both days, all subjects were shown sample stimuli and completed a series of practice sessions.

Practice sessions consisted of sets of 8 trials at longer SOAs (Experiment 1: 300ms x3, 240ms x3, 180ms x2; Experiments 2,3: 750ms x3, 600ms x3, 500ms x2). Subjects received feedback at the end of each set, and were required to achieve at least 7 correct trials in two consecutive practice sets before moving on to the training. Subjects were given as many opportunities as necessary to reach this level of performance, so long as there was enough time, although in Experiments 1 and 2, subjects who failed to achieve this in 20 sets were omitted from the study. In Experiment 3, subjects were required to reach this level of performance within 5 sets, and those who failed to achieve this in 5 sets were omitted from the study. This latter condition was run to address the possibility that the extra practice may have influenced the resultant pattern of learning.

Statistical Methods –We define learning for subjects in both the A-AB and AB-AB conditions as the difference in performance between the A_1 and A_2 sessions in each experiment. However, looking at transfer in the TDT required us to compare between different sessions depending on condition. In the A-AB groups transfer is the relationship between the A_1 and the B_2 sessions (e.g on different days), while for the AB-AB groups this is between the A_1 and the B_1 sessions (e.g. both on day 1). In both cases we are comparing the first presentation of the A stimuli to the first presentation of the B stimuli. However, we note that the magnitude of transfer, and perhaps the mechanisms, cannot be directly compared between these conditions. Statistical analyses of learning and transfer are conducted using between-subject comparisons of these within-subject difference scores. Additionally, although learning is defined in the same way in both the A-AB and the AB-AB conditions, because we need to consider transfer separately between conditions we analyze both learning and transfer separately between conditions in our discussions of visual attention and gaming. The differences in the way transfer is defined in each condition are a consequence of the experimental design and therefore separating learning and transfer by condition for all experiments is necessary regardless of our results suggesting a lack of retroactive interference in our results.

Subjects in each experiment were divided into "high" and "low" visual attention skill (HiVA vs. LoVA; see Figure 1.3) groups based on a median split of the average of the estimate of subjects' thresholds for each of the 6.67° and 20° rings. The thresholds for each eccentricity was derived from the average of independent staircases, and the average of those two thresholds were used as selection criteria. This value was also used to determine the correlation between visual attention skill and video game experience.

Subjects were categorized into three groups based on their video game playing habits. Those who played at least 5 hours per week of first-person shooter or action games for the past six months were considered "action video game players" (AVGP). Those who reported not playing any first-person shooter or action games in the past six

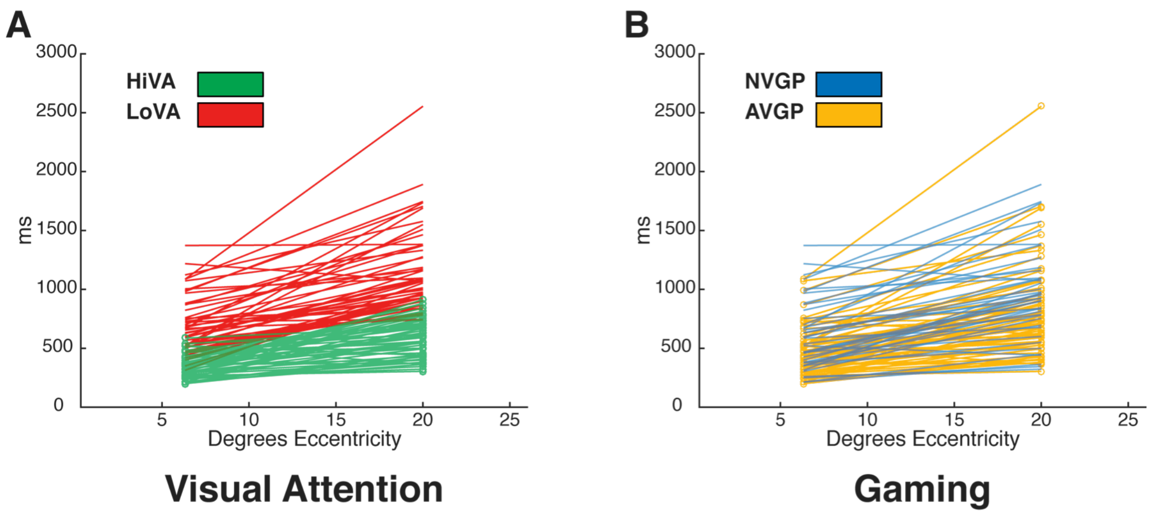


Figure 1.3. Performance in all experiments on the UFOV. Values shown are the calculated thresholds for the inner 6.67° and outer 20° target rings. Subjects are divided by their median (across experiments) performance on the UFOV (A; HiVA in green, LoVA in green) and by their video game experience (B; NVGP in blue, AVGP in yellow).

months were classified as "non video game players" (NVGP). Those who met neither criterion were classified as "in between" ("Tweeners"), who were left out of the main analyses. This method is the same as that used by Bavelier et al., (2011), where subjects self-report on video game experience. Only action video game play (defined narrowly as first-person-shooter and similar action games) is used as a criterion. Anecdotally, NVGP subjects usually play close to zero games, and subjects who play a variety of games (but not enough to qualify as AVGP) are most likely to appear as "Tweeners".

We conducted statistics on learning and transfer using both a proportion-correct based analysis and a threshold-based analysis. Our primary analyses involve accuracy on trials in which the subject responded correctly to the central stimulus, thus chance performance is 50%. We calculate baseline statistics (performance on the first training session on day 1) using the mean proportion-correct across all SOAs for that session, but for statistics on learning and transfer we look at the difference between the means for the two sessions in question.

To calculate thresholds, we fit the data with a psychometric function and calculated a 70% threshold for performance. We fit constrained psychometric functions on a session-by-session basis and used this as selection criterion for further analysis. Some subjects, particularly in Experiment 1, exhibited high variance in their performance leading to behaviorally implausible best-fits. For example, roving the order of difficulty levels in Experiment 1 resulted in a significant reduction of overall performance and many subjects failed to reach the threshold. Under other circumstances, a low baseline would result in rejecting subjects for failure to comply with the task. Since this is the result of our manipulation, we view the statistics based on proportion correct, which allows us to include all participants, to be more valid. Therefore, we constrained the function to exclude unrealistic behavior such as implying reduced performance at higher SOAs. From there, we used the MSE for each best fit to identify subjects whose behavior varied most significantly from the norm and further separated our participants into Psychometrically-Fit (PF) and Non-Psychometrically-Fit (NPF) groups. A detailed description of our selection method is be found in the Supplemental material. With few exceptions, the proportion-correct based analysis and the threshold-based analysis generate similar conclusions; see the Supplemental material for a complete outline of both analyses. Due to the agreement between the two analyses, for the main body of this article we present the proportion-correct results so as to include as many subjects as possible.

Results

Performance on the UFOV and relationship to Action Video Game Experience

Overall performance on the UFOV is comparable to findings previously reported in the literature (Dye, Hauser & Bavelier, 2009). Consistent with previous research we found a significant, but small, correlation between video game experience and threshold on the UFOV for the subject pool as a whole ($r = -0.23$; $p = 0.020$), however, this effect is relatively modest with less than 5% of the variance in performance on the UFOV task being explained by action video game experience. All subjects were included in this analysis, with no omissions based on subject performance. This weak correlation accounts for the mixed impact of visual attention and video game play on learning and transfer (discussed in later sections).

Moderating Effect of Roving vs. Sequential Difficulty Levels on Baseline TDT Performance

Subjects' overall performance is shown in Figure 1.4 for each experiment, session and training condition. Performance in the Sequential conditions (Experiments 2 and 3) on the initial training session (Day 1-A) is very consistent between Experiments 2 and 3 and is also consistent with previous research using these SOA values (Yotsumoto et al., 2009). In the Roving condition (Experiment 1), initial performance in this session was significantly impaired ($F_{1,1022} = 24.04$; $p < 0.001$). Despite these baseline differences between conditions, we found almost no interactions between this effect and the effects of attention skill and action video game experience on learning and transfer.

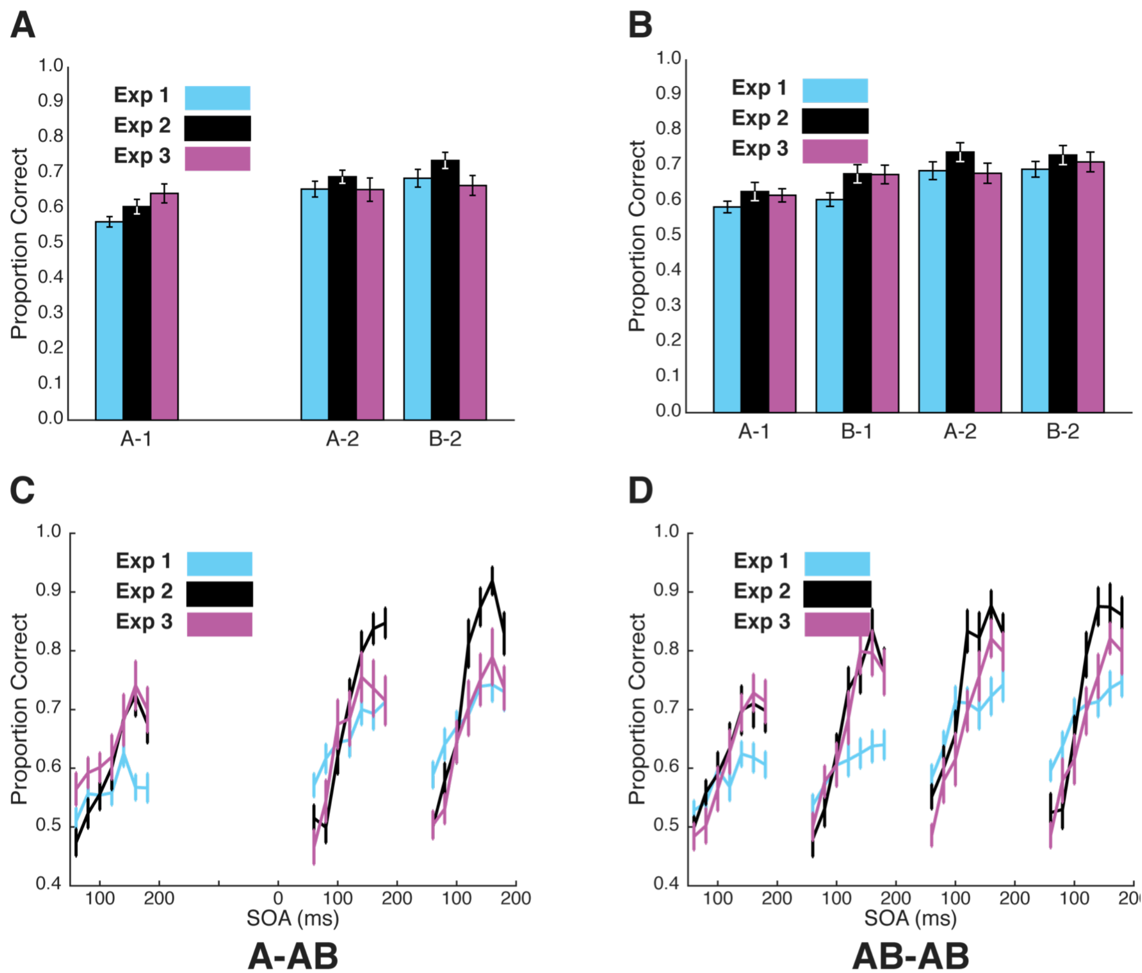


Figure 1.4. Performance in all three experiments (Exp 1: Cyan; Exp 2: Black; Exp 3: Purple) for each session of the TDT by mean across SOAs for each session (A, B) and by SOA (C,D) for A-AB (A, C) and AB-AB (B, D) conditions. Errorbars reflect standard error of the mean.

All subjects were included in this analysis, with no omissions based on subject performance. Therefore we present all experiments together in the remaining sections of the manuscript. Detailed breakdowns for each experiment, divided by both visual attention skill and video game experience may be found in the Supplemental Data.

Moderating Effects of Visual Attention

We first examined the extent to which visual attention skill (VA) as assessed by the UFOV moderated performance, learning and transfer in the TDT. For our comparison of visual attention we used the median split described above (HiVA vs. LoVA; see Figure 1.3). Across the three experiments there were 76 HiVA subjects and 78 LoVA subjects included in this analysis. All subjects were included in this analysis, subjects were selectively excluding from analysis only during the PF and NPF analyses (see supplemental material). Visual attention had a highly significant impact on baseline performance (see Figure 1.5 A,B) in session A_1 in this task ($F_{1,1022} = 39.31$, $p < 0.001$, proportion-correct statistics) which did not interact with other measures of interest. This effect was also found in threshold-based analysis, and the PF groups for both types of statistic, but was absent in the NPF groups. Moving forward, for brevity we will report the whole-experiment proportion-correct results alone, provided the other analyses produced the same results. See the Supplemental Material for complete statistical results. In the A-AB groups we found a highly significant effect of VA on the amount of learning ($F_{1,490} = 15.11$, $p < 0.001$) and transfer ($F_{1,490} = 10.58$, $p = 0.001$), although this did not appear in the threshold-based statistics. We saw a trend towards a significant effect for the AB-AB subjects, for both improved learning ($F_{1,504} = 3.52$, $p = 0.061$) and transfer ($F_{1,504} = 3.52$, $p = 0.061$), although our PF groups and the threshold-based analysis were not significant.

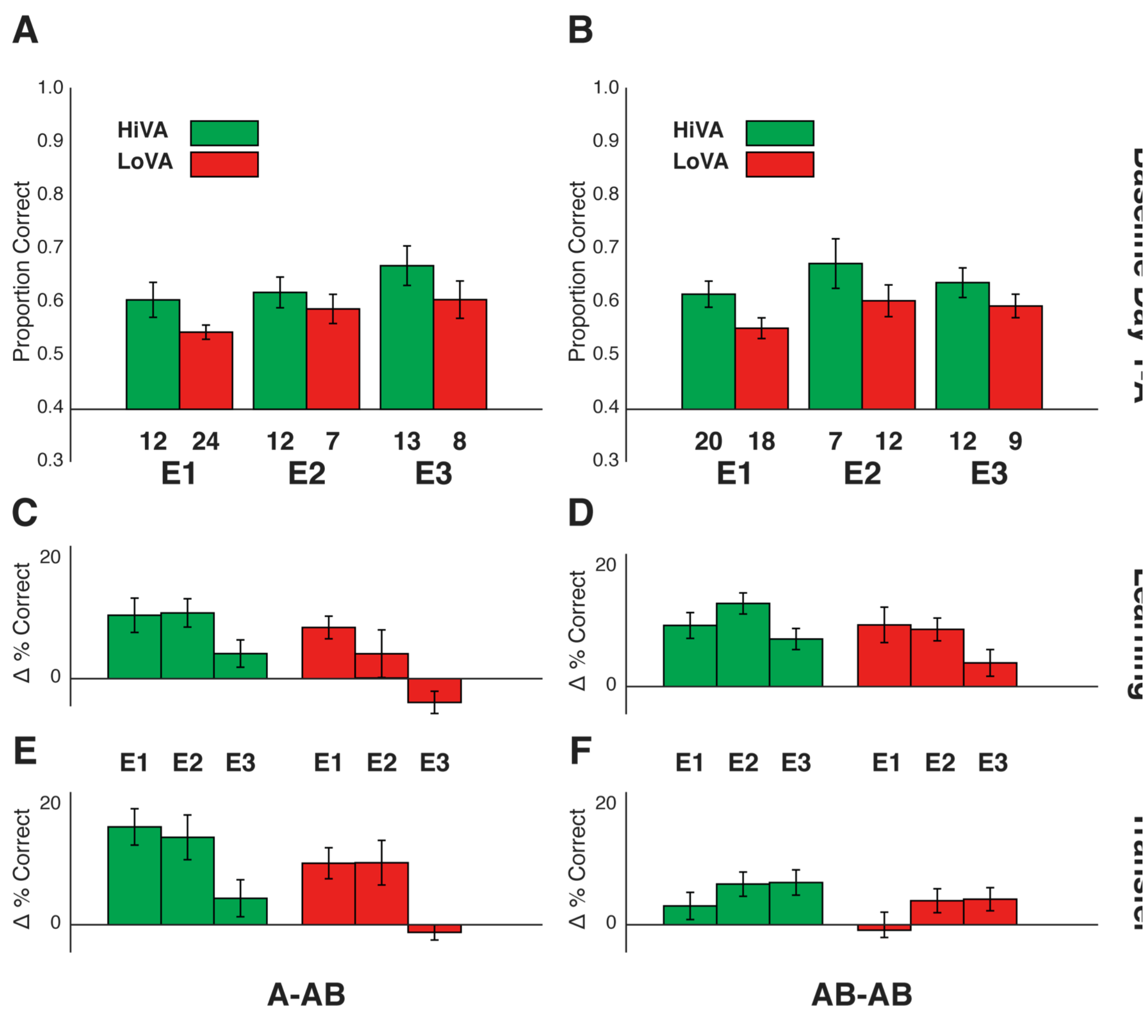


Figure 1.5. Effects of Visual Attention on learning and transfer. Baseline performance (session A_1) in all three experiments for A-AB (A) and AB-AB (B) conditions for proportion correct. Change in percent correct in learning (C, D) and transfer (E, F) in A-AB (C, E) and AB-AB (D, F) conditions. Subjects are divided by visual attention skill (VA), with HiVA (green) and LoVA (red) indicated separately. Errorbars reflect standard error of the mean. Group totals are listed below each column. Graphs reflect mean performance per session (A, B) and difference between mean performance on relevant sessions for comparison (C, D, E, F).

Interestingly, while VA had a significant impact on learning and transfer it did not interact significantly with other factors of interest like experimental condition or even trial difficulty in any of our analyses (Figure 1.5). These results suggest that existing attention abilities have a substantial effect on performance and a moderate effect on learning of the TDT task.

Moderating Effects of Action Video Game Experience

Next we looked at the effect of action video game play on performance, learning and transfer in the TDT. Across the three experiments there we have a large number of participants eligible for the analyses (41 NVGP subjects and 64 AVGP), however numbers of participants in individual conditions (noted in Figure 1.6) were more modest. Action video game play did not have a consistent or significant impact on baseline performance (see Figure 1.6 A,B) in session A_1 in this task ($F_{2,994}$ = 0.21, p = 0.810). In the A-AB groups we found significant effects of video game experience on the amount of learning ($F_{1,308}$ = 5.16, p = 0.026) but not on transfer ($F_{1,308}$ = 0.33, p = 0.565) but this effect on learning did not replicate in our PF groups. For AB-AB there were no significant effects on either learning ($F_{1,343}$ = 0.02, p = 0.882) or transfer ($F_{1,343}$ = 0.05, p = 0.826). Action video game experience did not interact with other factors of interest like experimental condition or even trial difficulty in any of our analyses (Figure 1.6). These results suggest that, although there is a statistically significant relationship between action video game play and attention, the relationship is not impactful enough to ensure that both factors display similar effects on learning and transfer in the TDT.

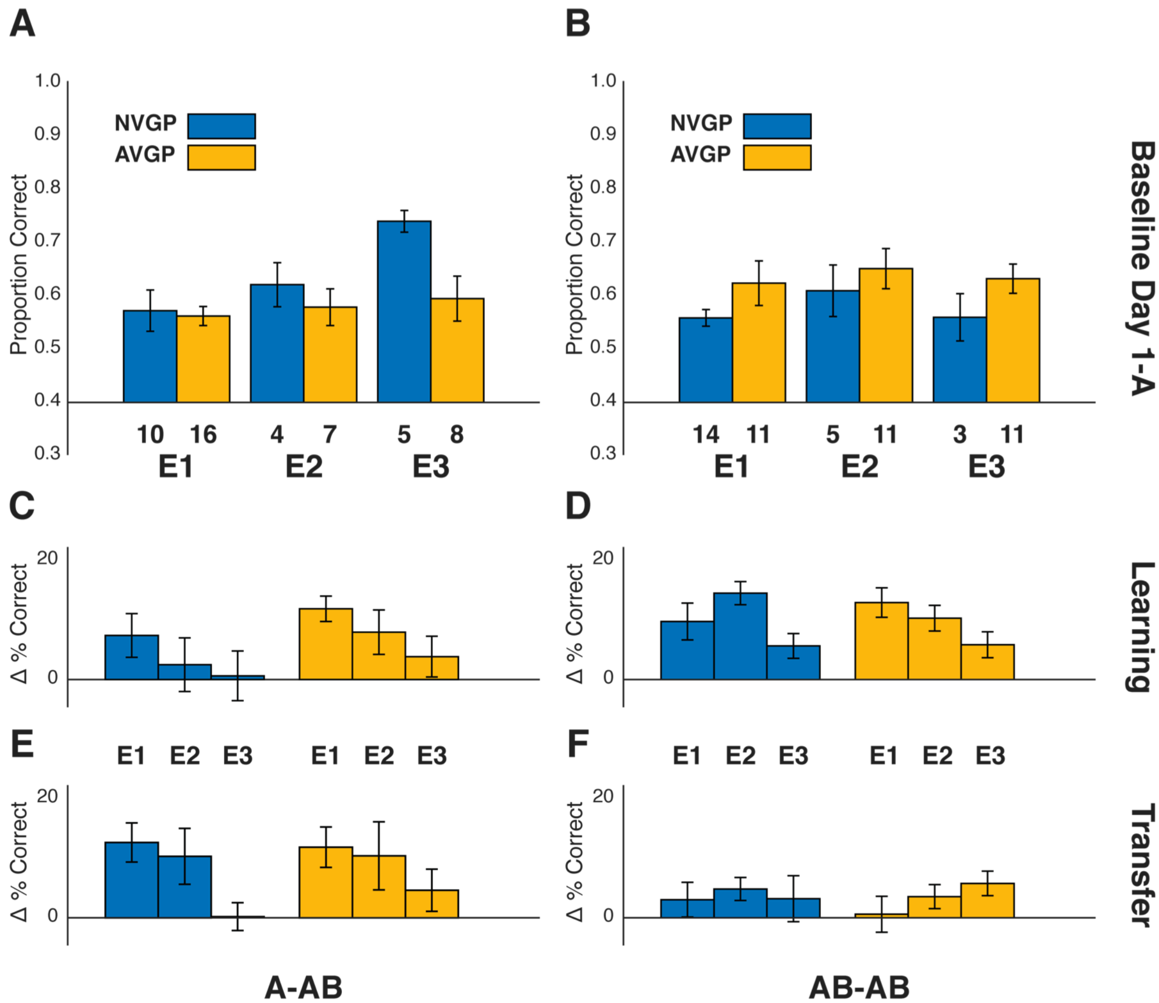


Figure 1.6. Effects of Action Video Game play on learning and transfer. Baseline performance (session A_1) in all three experiments for A-AB (A) and AB-AB (B) conditions for proportion correct. Change in percent correct in the learning (C, D) and transfer (E, F) in A-AB (C, E) and AB-AB (D, F) conditions. Subjects are divided by video game experience, with NVGP (blue) and AVGP (yellow) indicated separately. Errorbars reflect standard error of the mean. Group totals are listed below each column. Graphs reflect mean performance per session (A, B) and difference between mean performance on the relevant sessions for comparison (C, D, E, F).

Interactions between visual attention and action video game play on the TDT

Thus far we have treated each experiment separately and evaluated the effects of visual attention and of action video game independently from one another. Because separated subjects using a median split for the visual attention analyses, the two groups are approximately evenly divided between experiments. However, due to the self-report nature of video game experience some of the groups are unevenly distributed between experiment and condition. Therefore, we combined across experiments and conducted additional analyses of learning and transfer with these factors included together (See Figure 1.7). For learning, we found a significant main effect of gaming ($F_{1,322}$ = 5.12, p = 0.024), but not visual attention ($F_{1,322}$ = 2.38, p = 0.124) and no interaction between the two for our A-AB subjects.

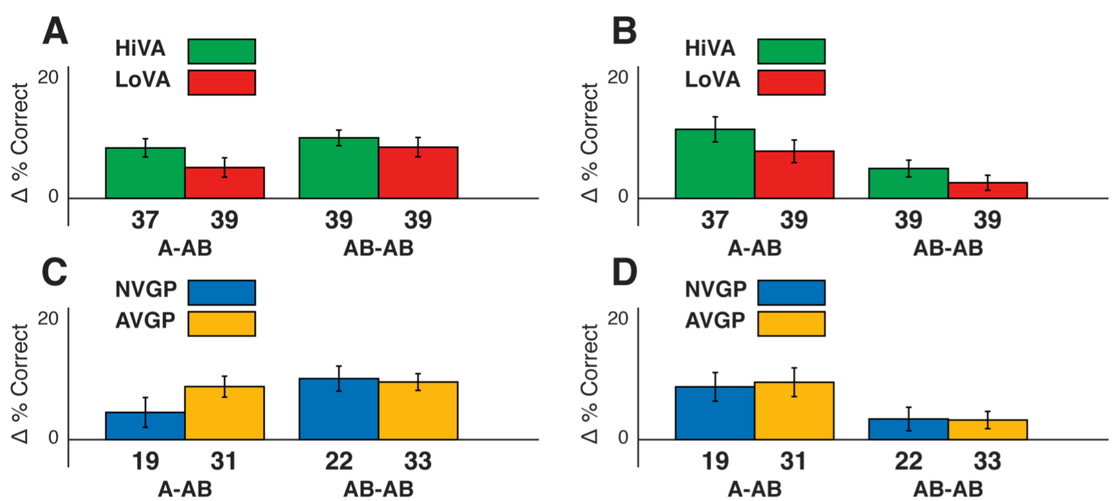


Figure 1.7. Effects of Visual Attention and Action Video Game Play on Learning and Transfer Across Experiments. Change in performance for learning (A, C) and transfer (B, D) in the TDT. Errorbars reflect standard error of the mean. Group totals are listed below each column. Graphs reflect the difference between mean performance on the relevant sessions for each comparison.

For the AB-AB subjects there were no significant effects, but there was a trend towards a main effect of visual attention ($F_{1,357}$ = 2.59, p = 0.108). For transfer, we found no significant effects for either the A-AB or AB-AB subjects, although there was a trend for a main effect of visual attention in the AB-AB group ($F_{1,357} = 2.75$, $p = 0.098$). These results again fail to support a consistent relationship between video game play and learning on the TDT or a consistent interaction between video game play and attention in moderating learning and transfer effects.

Moderating effects of interference on the TDT

We next examined whether the presence of the second training session in the AB-AB condition led to interference in learning for the A condition, as has been found in prior research (Yotsumoto et al., 2009; McDevitt et al., 2015) and whether visual attention skill or video game play moderated such interference. As we can see in Figure 1.8, we failed to find evidence of interference in any of the experiments. We conducted this analysis twice, once with VA as a factor and again with video game play as a factor. In contrast to interference, we found the subjects in the AB-AB condition demonstrated learning regardless of whether we include either visual attention ($F_{1,994} = 11.94$, $p < 0.001$) and video game play ($F_{1,651} = 9.52$, $p = 0.002$; These refer to the main effect of condition for the whole subject population; note this value excludes "Tweeners"). There

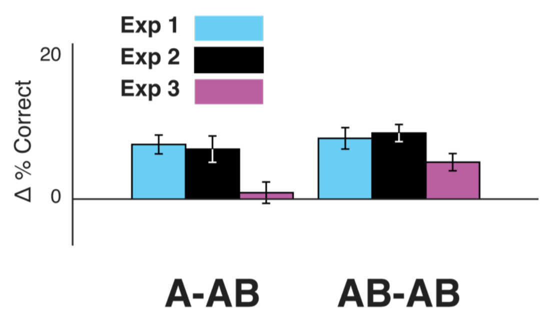


Figure 1.8. Interference in TDT. Learning for each experimental condition for all three experiments (Exp 1: Blue; Exp 2: Black; Exp 3: Purple). Errorbars reflect standard error of the mean. Graphs reflect the difference between mean performance on the relevant sessions for comparison.

is a trend towards an interaction between condition and video game play ($F_{1,651} = 3.04$, $p = 0.082$), however, while we would expect the presence of an interference condition to reduce learning, as we can see in Figure 1.6, we actually found that learning was improved for the NVGP in the AB-AB. Condition does not interact significantly with any of the other factors in these analyses. These results suggest that interference between background textures is not a ubiquitous effect of TDT training, and that it isn't easily predicted by attention skill or action video experience. Further, this effect is not the result of floor effects, as this result holds true in the PF/NPF analysis.

Discussion

The purpose of this study was to systematically examine how individual differences may moderate, and how differences in training task structure may mediate learning and transfer in perceptual learning. Results showed a significant effect of visual attention on all measures of interest but an inconsistent relationship between action video game play on performance and learning for the TDT. In terms of the distinct aspects of the task itself, randomizing the order of trial difficulties has a significant impact on participants' overall performance on the task. Finally, we were unable to replicate interference between background orientations in the TDT, and in some cases found substantial learning in the AB-AB condition comparable to the non-interference A-AB condition. Together, these results show that both individual and task factors impact performance on the TDT, however, not all relationships previously reported in the literature are simply replicated.

Our most robust finding was that visual attention skill had a significant effect on all aspects of performance on the TDT. Improved ability to allocate visual attention resulted in better baseline performance, enhanced learning, and improved transfer. The TDT is a task that relies upon split attention, and so it was unsurprising that individuals with superior skills in this area showed superior performance. This is consistent with research showing that attention can play a key role in mediating perceptual learning (Szpiro & Carrasco, 2015; Donovan, Szpiro & Carrasco, 2015; Schoups et al., 2001). Of note, there are multiple aspects of attention and even visual attention which influence perceptual learning that may be involved in the texture discrimination task. Wang et al. (2013) found that much of TDT learning can be attributed to temporal learning by either narrowing the temporal window of attention or increasing the speed of processing. Therefore it may be the case that if we had measured visual attention using a different task which relies less on rapid temporal processing we would draw different conclusions on the relationship between VA and learning.

The finding that action video game play was at best inconsistently related to performance on the TDT is unexpected. Action video game play has been shown to have wide-ranging effects on a wide variety of tasks (Green & Bavelier, 2012; Bejjanki et al., 2014). However, we found only a weak relationship between action video game play and the

UFOV and performance and learning on the TDT task. These findings are in contrast to recent research using the TDT that has found that video game experience (although not action video games) has a significant impact on performance (Kim et al., 2015). Furthermore, it is thought that extensive experience with action video game play can facilitate learning (so-called “learning to learn”; Bavelier et al., 2012). However, because action video game play was not correlated with UFOV performance in this sample (and visual attention *was* strongly related to performance in the TDT) this may help explain why we find little relationship between action video game play and learning in this population. It is unclear whether this is due to cohort changes in the use of technology, including mixed games that may not be considered action video games,(Dale & Green, 2017) or whether other factors are at play here. Including a larger sample with multiple study locations may be required to reconcile the extent to which these results are valid to other populations.

Previous research (Seitz et al., 2005) suggests that retrograde interference can be induced by immediately training a second task condition after an initial training session, causing a disruption of learning for the first-trained condition. This was initially found using a hyperacuity task (Seitz, et al., 2005) as well as with contrast stimuli (Adini et al., 2004) and the TDT (Yotsumoto et al., 2009; McDevitt, Duggan & Mednick, 2015). However, findings of retrograde interference are inconsistent in the PL literature, for example Aberg and Herzog (2010) failed to find retrograde interference in a number of training conditions including an attempt to replicate Seitz et al., (2005); although Hung and Seitz

(2011) replicated both studies (Seitz et al., 2005 and Aberg & Herzog; 2010) and suggested the discrepancy between them as being related to differences in eye-movements across the studies. In the case of the TDT it has been found that interference can be rescued through REM sleep (McDevitt, Duggan & Mednick, 2015) suggesting that factors like sleep can mediate findings of interference. Alternatively, it is possible that the use of the UFOV task prior to TDT training may have led to an increased transfer effect, which would be consistent with Wang et al.'s (2004) findings of the relative importance of temporal learning on the TDT. The UFOV task requires focused attention on briefly-presented stimuli, much like the TDT. And while there are significant differences such as the unknown location of the target in the UFOV, there may be enough similarities to stimulate the effects we found in this study. Further, while we attempted to carefully replicate the conditions of Yotsumoto et al. (2009), some parameters still differed; for example we used a 20ms target duration whereas their targets were present for only 13ms. Future investigations will need to understand what factors make the key difference in these tasks. This is further complicated by the fact that many of our subjects were unable to reach the same standards of performance as those in Yotsumoto et al.'s (2009) study. This reduction in overall performance may be responsible for the differences we see in retrograde interference between these studies. Again, while we cannot claim that retrograde interference does not occur in the TDT, our results do suggest that it is not ubiquitous and that further research is required to understand the conditions where it may or may not be observed.

In spite of the extensive sample size and replication of previously-used task parameters there are still several important limitations on these results. In particular, participants in our sample found the trial difficulties (SOAs) to be exceptionally difficult. In Experiments 1 and 2 subjects had ample practice sessions to prepare themselves for the task, but some required many repetitions in order to reach adequate performance. This previous experience may have had a deleterious impact on their overall performance and learning. We conducted Experiment 3, which limited practice sessions, to address the potential impact of practice sessions and found that the addition of extra practice trials had a limited effect on the overall pattern of results. However, details of practice sessions have been shown to play an important role in perceptual learning in other training contexts (Ahissar & Hochstein, 1997; Seitz, Nanez, et al., 2005).

We also note that performance was poor in a number of participants that this may impact the overall pattern of results. First, this can have a direct influence on learning, as a number of studies show that difficulty during training can have a significant impact on perceptual learning (Hung and Seitz, 2014; Weinliang and Seitz, 2018; Ahissar & Hochstein, 1997), and has been argued to potentially give rise to a different profile of learning across the brain systems that underlie learning (Maniglia and Seitz, 2017). Further, this also has an influence on how learning is characterized in that this differs from how studies characterize learning in the TDT (Karni & Sagi, 1993; Harris & Sagi, 2015), however, it is consistent with other studies (e.g. Yotsumoto et al., 2009). While many results are consistent across methods of analyses, some are not. We give a detailed presentation of the results in the supplemental information so that readers can come to their own conclusions regarding the impact of the different approaches to analysis.

In conclusion, we find that individual differences and training variants play a key role in shaping perceptual learning. While factors such as attention skill and how trials of different difficulties are interleaved have substantial impact on the TDT, other factors such as video game experience and block-wise interference showed less consistent impact on performance and learning. It is also likely that differences in performance levels experienced during training may moderate these effects (CITE Wenliang and Seitz, J. Neurosci 2018). These results challenge previous findings in the literature and suggest that larger sample sizes, and clearer understanding of subject and task factors is required to better understand how, and why, effects of perceptual learning differ across studies.

References

Aberg, K. C., & Herzog, M. H. (2010). Does perceptual learning suffer from retrograde interference? *PloS One*, *5*(12), e14161.

Adini, Y., Wilkonsky, A., Haspel, R., Tsodyks, M., & Sagi, D. (2004). Perceptual learning in contrast discrimination : The effect of contrast uncertainty, 993–1005. doi:10.1167/4.12.2

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

Ball, K., Beard, B. L., Roenker, D. L., Miller, R. L., & Griggs, D. S. (1988). Age and visual Search: expanding the useful field of view. *Journal of the Optical Society of America, A*, *5*(10), 2210–2219.

Bavelier, D., Achtman, R. A., Mani, M., & Foecker, J. (2011). Neural bases of selective attention in action video game players. *Vision Research*.

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.

Bays, B. C., Visscher, K. M., Le Dantec, C. C., & Seitz, A. R. (2015). Alpha-band EEG activity in perceptual learning. *Journal of Vision*, *15*(10), 7-7.

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.

Berard, A. V., Cain, M. S., Watanabe, T., & Sasaki, Y. (2015). Frequent video game players resist perceptual interference. *PloS one*, *10*(3), e0120011.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436.

Brashers-Krug, T., Shadmehr, R., & Bizzi, E. (1996). Consolidation in human motor memory. *Nature*, *382*(6588), 252–255.

Censor, N., Bonneh, Y., Arieli, A., & Sagi, D. (2009) Early-vision brain responses which predict human visual segmentation and learning. *Journal of Vision*, *9*(4):12, 1-9.

Censor, N., Karni, A., & Sagi, D. (2006). A link between perceptual learning, adaptation and sleep. *Vision Research*, *46*(23), 4071-4074.

Dale, G., & Green, C. S. (2017). The changing face of video games and video gamers: Future directions in the scientific study of video game play and cognitive performance. *Journal of Cognitive Enhancement*, *1*(3), 280-294.

Dobres, J., & Seitz, A. R. (2010). Perceptual learning of oriented gratings as revealed by classification images. *Journal of Vision*, *10*(13), 8-8.

Donovan, I., Szpiro, S., & Carrasco, M. (2015). Exogenous attention facilitates location transfer of perceptual learning. *Journal of Vision*, *15*(10), 11-11.

Dye, M. W., Hauser, P. C., & Bavelier, D. (2009). Is visual selective attention in deaf individuals enhanced or deficient? The case of the useful field of view. *PloS one*, *4*(5), e5640.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*(6939), 534–7.

Green, C. S., & Bavelier, D. (2012). Learning, attentional control and action video games. *Current Biology*, *22*(7), R197-R206.

Gu, Y., Liu, S., Fetsch, C. R., Yang, Y., Fok, S., Sunkara, A., DeAngelis, G. C. & Angelaki, D. E. (2011). Perceptual learning reduces interneuronal correlations in macaque visual cortex. *Neuron*, *71*(4), 750-761.

Harris, H., Gliksberg, M., & Sagi, D. (2012). Generalized perceptual learning in the absence of sensory adaptation. *Current Biology*, *22*(19), 1813-1817.

Harris, H., & Sagi, D. (2015). Effects of spatiotemporal consistencies on visual learning dynamics and transfer. *Vision Research*, *109*, 77-86.

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

Hung, S.-C., & Seitz, A. R. (2011). Retrograde interference in perceptual learning of a peripheral hyperacuity task. *PloS One*, *6*(9), e24556.

Hung, S. C., & Seitz, A. R. (2014). Prolonged training at threshold promotes robust retinotopic specificity in perceptual learning. *Journal of Neuroscience*, *34*(25), 8423-8431.

Huang, J., Liang, J., Zhou, Y., & Liu, Z. (2017). Transfer in motion discrimination learning was no greater in double training than in single training. *Journal of Vision*, *17*(6), 7-7.

Karni, A, & Sagi, D. (1991). Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(11), 4966–70.

Karni, A. & Sagi, D. (1993). The time course of learning a visual skill. *Nature*, 365, 250-252.

Kim, Y. H., Kang, D. W., Kim, D., Kim, H. J., Sasaki, Y., & Watanabe, T. (2015). Real-time strategy video game experience and visual perceptual learning. *Journal of Neuroscience*, *35*(29), 10485-10492.

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.

Liang, J., Zhou, Y., Fahle, M., & Liu, Z. (2015). Limited transfer of long-term motion perceptual learning with double training. *Journal of Vision*, *15*(10), 1-1.

Maniglia, M. & Seitz, A. R. (2018). Towards a whole brain model of perceptual learning. *Current Opinion in Behavioral Sciences*, 20, 47-55.

McDevitt, E. A., Duggan, K. A., & Mednick, S. C. (2015). REM sleep rescues learning from interference. *Neurobiology of Learning and Memory*, *122*, 51-62.

Mednick, S., Nakayama, K., & Stickgold, R. (2003). Sleep-dependent learning: a nap is as good as a night. *Nature Neuroscience*, *6*(7), 697–8.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437–442.

Sagi, D. (2011). Perceptual learning in vision research. *Vision Research*, *51*(13), 1552-1566.

Schoups, A., Vogels, R., Qian, N., & Orban, G. (2001). Practising orientation identification improves orientation coding in V1 neurons. *Nature*, *412*(6846), 549.

Seitz, Nanez, Holloway, Koyama, and Watanabe (2005), “Seeing what isn’t there; the costs of perceptual learning”, *PNAS*, Jun 21;102(25):9080-5

Seitz, A. R., & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive Sciences*, *9*(7), 329–334.

Seitz, A. R., Yamagishi, N., Werner, B., Goda, N., Kawato, M., & Watanabe, T. (2005). Task-specific disruption of perceptual learning. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(41), 14895–900.

Szpiro, S. F., & Carrasco, M. (2015). Exogenous attention enables perceptual learning. *Psychological Science*, *26*(12), 1854-1862.

Wang, R., Cong, L. J., & Yu, C. (2013). The classical TDT perceptual learning is mostly temporal learning. *Journal of Vision*, *13*(5), 9-9.

Li K. Wenliang, L. K. & Seitz, A. R. (2018). Deep neural networks for modeling visual perceptual learning. *Journal of Neuroscience*, *38*(27), 6028-6044.

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.

Yotsumoto, Y., Chang, L. H., Watanabe, T., & Sasaki, Y. (2009). Interference and feature specificity in visual perceptual learning. *Vision Research*, *49*(21), 2611–2623.

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), 4-4.

Zhang, J. Y., & Yu, C. (2016). The transfer of motion direction learning to an opposite direction enabled by double training: A reply to Liang et al.(2015). *Journal of Vision*, *16*(3), 29-29.

Using a Joint Criterion for

Participant Selection in Perceptual Learning Methods

Theodore Jacques[1] & Aaron R. Seitz[1]

Author Affiliations:

[1]Department of Psychology, University of California, Riverside, Riverside, CA 92521

Correspondence should be addressed to:

Theodore Jacques

theodore.jacques@gmail.com

(567)-203-8232

Abstract

The possibility of “stimulus independent lapses” in attention by participants during perceptual learning training tasks can cause significant noise in data collection and interpretation. These lapses are typically resolved by excluding subjects who exhibit poor performance. As part of an existing study investigating moderating effects on learning and transfer in a Texture Discrimination Task we found reduced performance was due in part to one of the experimental manipulations. Here we discuss a novel approach for accounting for data influenced by these lapses independent from baseline performance scores. Then we discuss the impact of these subsequent analyses on the conclusions drawn based on the whole-dataset analyses. We found that our initial conclusions were supported by these latter analyses.

Introduction

This chapter deals with the novel data analytic methods we developed in order to conduct a follow-up analysis for the Texture Discrimination experiment described in detail in Chapter 1. Much of this work was inspired by one comment we received from a reviewer during the first round of peer review for that manuscript. To begin the chapter, we will briefly review the circumstances of the question and the line of reasoning which ultimately led to such a substantial development of novel techniques. Next we will describe those novel approaches in detail, and finally conclude the chapter with a discussion of the empirical results they produced relative to the perceptual experiment in Chapter 1.

The reviewer comment which inspired this work was focused on the type of data we selected to report our results. In Chapter 1 we present and discuss results almost exclusively based on proportion-correct. This is the most straightforward way to present Texture Discrimination data, since it requires no intermediate steps to transform the results. However, this is not the standard method of reporting in the field. Typically, the raw proportion-correct results are fit with a psychometric function and the stimulus intensity corresponding to the intercept between the function and a particular performance level (the threshold) is used instead (see Karni & Sagi, 1991; Ahissar & Hochstein, 1997; Stickgold, James & Hobson, 2000, and many more besides). Both methods are derived from the same numbers and reflect the same information (and indeed can be seen on the same figure) and roughly correspond to the difference between focusing on the x-axis or the y-axis of a graph. The reviewer asked why we chose to report performance in terms of proportion-correct, when reporting a threshold is the standard in the field.

In the process of fitting the data with a psychometric function and calculating the threshold we had stumbled into a numerical problem. Psychophysical data often takes the form of a particular proportion correct at each of several difficulty levels and the Texture Discrimination Task (TDT) is no exception. In the TDT the target stimulus is followed by the presentation of a visual mask and the amount of time before the mask appears (the Stimulus Onset Asynchrony (SOA)) defines the level of difficulty for the trial. Longer SOAs are easier and shorter SOAs are harder. Logically, fitting a psychometric function to report a 70% threshold necessitates that performance values on the measured SOAs should include this level of performance. In our experiment we duplicated the stimulus intensities used in a previous TDT article by Yotsumoto et al. (2009) and so we expected our chosen SOAs would span the sensitivity range for our subjects. However, upon conducting the study we were confronted by the fact that many of our subjects had not reached this level of performance even on the longest SOAs in our study. Furthermore, we understood that one of our experimental manipulations (roving trial difficulty, as discussed in Chapter 1) was causing this reduction in performance and therefore these low levels of performance could not be summarily dismissed.

Identifying the cause of reduced performance is critical in perceptual learning, and most often reduced performance is attributed to subject error. Wichmann & Hill (2001) refer to this type of subject error as "stimulus-independent *lapses*" and significant effort has gone into understanding their influence on the subsequent computation of the psychometric function (Harvey, 1986; Swanson & Birch, 1992). Perceptual tasks are often dull and repetitive, and so maintaining focus on task is a major concern in perceptual learning. These “lapses” are a polite way to talk about those moments when subjects simply aren’t paying attention like they should. When a subject is not properly focused on the task, their responses will be effectively random and their level of performance will drop towards chance. In truth there are many possible reasons for why a particular subject might show lower performance than others, but the fact that reduced performance also interferes with the ability to compute a threshold has given researchers a convenient excuse to resolve both problems at once. Typically, any subject with low overall performance is assumed to have made many “lapses” throughout the experiment, and their data is discarded on the basis that they did not follow instructions. This is frequently done even if there may be other factors contributing to reduced performance, and typically the factors at play are not explored in detail in favor of simply excluding participants. While this approach may be common in Texture Discrimination research and similar perceptual learning tasks, in this experiment we were unable to follow suit. As discussed in Chapter 1, we had found that our subjects showed reduced performance due to one of our manipulations. Therefore, we knew that we could not simply discard these subjects as “noise” in our data and that doing so would eliminate critical information relevant to our research questions. Ultimately we did discard these subjects and attempt to re-do our analyses to try and appease the reviewer (these results will be discussed later), however this line of reasoning revealed a new problem we could not ignore.

We could not discard low-performing subjects on the basis that they were "bad" subjects who did not follow instructions, but deliberating on the issue had highlighted the likelihood that our data set probably did contain "bad" subjects nonetheless. Could we be sure that the results we were reporting were not influenced by data from subjects who were inattentive or otherwise not participating to their best ability? The method to address this question is theoretically straightforward: if we can identify "bad" subjects and separate them from our "good" subjects, then repeating the analyses on each group separately will reveal the true results. If the effects we observed in the whole-subjects analysis appear in both subsequent analyses we can conclude that the effects are valid independent from subject "goodness" or "badness", and if they are visible in the "good" group but not the "bad" group then we can still say with some confidence that our interpretations are accurate. We were able to eventually separate our subjects in this way, and it was this process which includes our novel data analytic techniques. In the next section we will outline the numerical and logical methods used, and finally we will discuss the results of our follow-up data analysis.

Methods

Our ultimate objective in order to understand the results from Chapter 1 was to separate "good" subjects from "bad" subjects on the basis of a single criterion. As discussed earlier, we could not use mean performance on a given session as this criterion despite the fact it is the most frequent metric chosen for this purpose. However, during the course of fitting the data with a psychometric function we realized that the mean square error for our fits, usually a measure of goodness-of-fit, represented a viable alternative criterion.

Psychometric Function Fitting

We fit psychometric functions using numerical least-squares methods in Matlab (Mathworks, Natick, MA) and discovered that in some cases the best-fitting function resulted in behaviorally-implausible predictions. For example, we would expect subjects to do better on easier trials and so the slope parameter for a Logistic function (a commonly-used psychometric function) should be positive. However, in rare cases we found that the best-fitting slope parameter was negative or approximately zero, suggesting our subjects performed better on shorter SOAs and more poorly on longer SOAs. This is logically nonsense, and so we proceeded to constrain the psychometric functions to only “sensible” values to ensure behavioral plausibility in the resulting fit. By itself constraining the function is not particularly novel or unique. However, we recognized that while the mean squared error term reflects goodness-of-fit in the context of an unconstrained model, in the context of a constrained model the appropriate interpretation of this term changes and it becomes useful in this application to identify “bad” or atypically-behaving subjects.

The mean squared error (MSE; see D'Agostino, 1986 for additional resources on this subject) is frequently used as a goodness-of-fit measure. To calculate the MSE, find the shortest distance between each data point and the fitted function (the error of the fit), then square the distances. Finish by taking the mean for all data points in the set, hence the mean-squared-error. Logically, a good fit will describe a curve that lies close to each datapoint and so a good fit will have a low MSE. We used the lowest MSE from three different functions to select the best one for use later, and this is the way this value is used most frequently. However, due to the fact that we are using constrained models and the MSE takes a different meaning, we can apply the logic in reverse to determine which subjects are "good" and which are "bad",

Recall that we used functions which were constrained to behaviorally-plausible parameters, thus the MSE for the constrained models actually corresponds to the degree to which a subject's data resembles normal human task behaviors. If the subject's data has a particularly high MSE, it suggests that that subject is performing randomly relative to a normal behavioral pattern. Intermittent lapses in attention or other task-unrelated changes in behavior will result in a more randomly distributed pattern of responses and the measured results will differ significantly from the smooth pattern of a focused and attentive subject. Therefore, the correct interpretation of a comparatively high MSE in the context of a constrained model is as an indication of random, unfocused behavior. By first selecting for the lowest overall MSE to determine the best fitting function, and then selecting for the highest MSEs for all subjects in a session we were able to identify those subjects deviating most from expected patterns of behavior – our "bad" subjects.

To this end, we fit each individual session using three functions, a constrained Weibull function, a constrained logistic function, and an unconstrained linear function. Subjects in the A-AB group had each of their 3 sessions fit individually, and likewise the AB-AB group had each of their 4 sessions treated separately. The Weibull fit was constrained to a lapse rate between 0 and 1, a location parameter between the maximum and minimum SOAs for the experiment (180 and 60ms), and a positive slope parameter (0 to +inf). The Logistic fit was constrained to force upper and lower asymptotes between 0 and 1, a slope parameter between 0 and 1, and a location parameter between the maximum and minimum SOAs. The linear fit was unconstrained, and was included to account for particularly poor fits of the primary two types. Each fit was accompanied by a corresponding MSE, which was used in three ways. First, to identify which of the three functions provided the best fit, second as our metric of "goodness" and "badness", and third if the function for that subject intersected the 70% threshold we included that subject in subsequent threshold-based analyses (with one caveat, discussed below).

Psychometrically-Fit (PF) Subject Selection

One additional step is necessary before we can finally divide our participants into groups. These groups are derived from the degree to which subjects correspond to a typical psychometric function, so henceforth we will refer to them as Psychometrically-Fit (PF) and Non-Psychometrically-Fit (NPF) groups. In Chapter 1 we describe a perceptual learning experiment which investigated measures of learning and transfer in the TDT. Therefore, we needed to consider subjects not simply on the basis of a single session, but on the basis of two sessions together. For the Learning-PF group, this corresponded to the A-sessions on days 1 and 2. For the Transfer-PF group the sessions involved differed based on the experimental condition of each subject (see Chapter 1 for a more detailed discussion of this aspect of the experiment).

Our goal was to divide the subjects into two groups of approximately equal size. For dividing subjects into groups we considered only the best-fitting function for each session independently, even if this meant mixing-and-matching different fitting methods within any given subject. On a session-by-session basis we identified the best ~70% of subjects based on the lowest $\mu + \frac{1}{2}\sigma$ of mean squared error. PF-Learning and PF-Transfer subjects were required to meet this criterion for both sessions of the comparison being made. Thus we are not simply in need of a single cutoff criterion, but a joint-criterion for two sessions. Subjects in the threshold-based statistics section of our results also needed to meet this joint criterion for both sessions, this is the caveat we mentioned above.

The MSE cutoff of half of a standard-deviation above the mean (for that session) was chosen due to an interesting mathematical quirk. For a normal distribution, $\mu + \frac{1}{2}\sigma$ accounts for approximately 70% of the sample population. And if the two sessions were independent, then the probability of a subject falling in both categories would be 70%-squared, or about 49%. We know that the distribution of MSEs per session is not in fact normal and that two sessions for the same subject cannot possibly be considered mathematically independent, however for the purposes of dividing our subjects into two groups of approximately equal size this is a useful standard for a joint-cutoff criterion. To our knowledge no articles have previously reported a joint criterion, so we have no alternative algorithm for comparison. Although we have not conducted extensive statistical tests or replications to validate this approach for formal recommendation, it is clear from the group totals that this method is adequate for the desired goal of an approximately-median split between subject groups. See Table 2.1 for a complete list of the total subjects in each group of our subsequent data analysis, including breakdowns by visual attention and action video game experience (as discussed in Chapter 1).

		Whole Experiment	PF-Learning	PF-Transfer	NPF-Learning	NPF-Transfer
Experiment 1	Total Subjects	74	41	39	33	35
A-AB	HiVA	12	7	6	5	6
	LoVA	24	10	11	14	13
	NVGP	10	4	4	6	6
	AVGP	16	9	7	7	9
AB-AB	HiVA	20	12	10	8	10
	LoVA	18	12	12	6	6
	NVGP	14	7	10	7	4
	AVGP	11	9	6	2	5
Experiment 2	Total Subjects	38	26	26	12	12
A-AB	HiVA	12	9	8	3	4
	LoVA	7	5	4	2	3
	NVGP	4	3	3	1	1
	AVGP	7	5	3	2	4
AB-AB	HiVA	7	6	6	1	1
	LoVA	12	6	8	6	4
	NVGP	5	4	3	1	2
	AVGP	11	7	8	4	3
Experiment 3	Total Subjects	42	28	22	14	20
A-AB	HiVA	13	9	8	4	5
	LoVA	8	4	3	4	5
	NVGP	5	4	2	1	3
	AVGP	8	5	6	3	2
AB-AB	HiVA	12	8	6	4	6
	LoVA	9	7	5	2	4
	NVGP	3	2	2	1	1
	AVGP	11	9	6	2	5

Table 2.1. Subject breakdown for PF and NPF groups by experiment and experimental condition. Note that NVGP and AVGP totals will not sum to the experiment total due to omission of the "in between" subjects in the video game analysis. These subjects are included in the visual attention totals.

Results

Proportion-Correct Analysis

Results in the proportion-correct analysis refer to the results in Figures 2.1-2.4. In these figures we include baseline Day 1-A performance for both all subjects, and learning and transfer results for only the relevant analyses. Statistical results for all groups, along with the whole-subject results (duplicated here from Chapter 1 for comparison) will be discussed individually in each section with reference to the appropriate figure.

Moderating Effect of Roving the Order of Trial Difficulty

Baseline performance for both PF and NPF groups demonstrate the same pattern as the whole subject pool, with baseline proportion correct being significantly lower in Experiment 1 (Figures 2.1-2.4; A, B, E and F). NPF groups showed a reduced, but still significant, effect for this pattern, indicating that the observed effect reduced performance due to roving the order of trial difficulty levels in Experiment 1 did not depend on the PF/NPF distinction. See Table 2.2 for a full list of these results.

	Main Effect of Experiment		partial-η^2	Experiment 1	Experiment 2	Experiment 3
All Subjects	$F_{1,1022}$ =	24.04***	0.0183	0.57(0.14)	0.62(0.17)	0.63(0.18)
PF-Learning	$F_{1,609}$ =	16.08***	0.0177	0.59(0.14)	0.63(0.17)	0.66(0.18)
PF-Transfer	$F_{1,553}$ =	17.47***	0.0201	0.59(0.14)	0.64(0.18)	0.66(0.18)
NPF-Learning	$F_{1,357}$ =	3.92*	0.0093	0.55(0.13)	0.58(0.15)	0.58(0.16)
NPF-Transfer	$F_{1,413}$ =	3.75†	0.0077	0.55(0.14)	0.57(0.14)	0.60(0.16)

Table 2.2. Main effect of experiment on baseline performance on Day 1-A session with mean proportion correct (with SD) for each experiment for whole sample, PF, and NPF groups.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; † $p < 0.10$

Moderating Effects of Visual Attention

The main effect of visual attention on baseline Day 1-A performance is statistically significant in all PF and NPF groups, although there is a clear reduction in the effect for our NPF groups. This indicates that the effect of visual attention on overall performance on the TDT did not depend on the PF/NPF distinction. The significant effect of VA on learning in the A-AB subjects was borne out in the PF-Learning and NPF-Learning groups, although the trend we saw in the AB-AB subjects was not present in the PF group (Figures 2.1 and 2.3). We include the statistics for the PF-Transfer and NPF-Transfer groups in the table, however it is difficult to interpret the unique effects of a transfer-selected group of subjects on learning. Therefore, while all these figures are included in the statistical tables we will omit them from discussion. We were surprised to see that the effect on transfer discussed in the whole-experiment analyses is not reflected in the transfer-PF group in either the A-AB or AB-AB comparisons and instead seems to be driven by subjects included in the Learning-NPF and Transfer-NPF groups. In Chapter 1 we concluded that the primary effect of interest from visual attention was on overall performance and learning, so these results, while puzzling, do not change our overall interpretation of the experiment as a whole. See Table 2.3 for a full list of these results.

Main Effect of Visual Attention on:

	Baseline Day 1-A Performance	partial-η^2		
All Subjects	$F_{1,1022}$ = 39.31***	0.0300		
PF-Learning	$F_{1,609}$ = 33.56***	0.0369		
PF-Transfer	$F_{1,553}$ = 38.82***	0.0447		
NPF-Learning	$F_{1,357}$ = 4.69*	0.0111		
NPF-Transfer	$F_{1,413}$ = 6.86**	0.0141		
	Learning : A-AB Subjects	**partial-η^2**	**Learning : AB-AB Subjects**	**partial-η^2**
All Subjects	$F_{1,490}$ = 15.11***	0.0254	$F_{1,504}$ = 3.52†	0.0064
PF-Learning	$F_{1,266}$ = 4.36*	0.0132	$F_{1,315}$ = 0.11	0.0003
PF-Transfer	$F_{1,238}$ = 2.78†	0.0090	$F_{1,287}$ = 0.09	0.0003
NPF-Learning	$F_{1,182}$ = 13.71***	0.0515	$F_{1,147}$ = 5.70*	0.0287
NPF-Transfer	$F_{1,210}$ = 14.74***	0.0528	$F_{1,175}$ = 7.99**	0.0354
	Transfer : A-AB Subjects	**partial-η^2**	**Transfer : AB-AB Subjects**	**partial-η^2**
All Subjects	$F_{1,490}$ = 10.58**	0.0176	$F_{1,504}$ = 3.16†	0.0064
PF-Learning	$F_{1,266}$ = 1.46	0.0043	$F_{1,315}$ = 1.24	0.0035
PF-Transfer	$F_{1,238}$ = 1.26	0.0037	$F_{1,287}$ = 0.10	0.0003
NPF-Learning	$F_{1,182}$ = 16.98***	0.0609	$F_{1,147}$ = 16.80***	0.0747
NPF-Transfer	$F_{1,210}$ = 8.75**	0.0339	[1] $F_{1,175}$ = 9.05**	0.0399

Table 2.3. Main effect of visual attention on baseline performance during Day 1-A session, learning, and transfer (split by experimental condition) for whole sample, PF, and NPF groups.

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.10
[1] This analysis indicated a nearly-significant SOA x VA interaction, p = 0.059.

Moderating Effects of Action Video Game Experience

The lack of a main effect of video game experience on baseline Day 1-A performance replicated in both PF groups. The significant effect of gaming on learning in the A-AB subjects was not present in the learning-selected PF group, although it was apparent in the NPF group. This is puzzling to interpret, as those subjects primarily driving the overall effect on learning seem to be those who are behaving most randomly. This suggests that the effect may be a statistical aberration, particularly in light of the absence of an effect in the AB-AB group of any kind. The lack of effect in the transfer comparisons was borne out in the PF groups in both conditions, in spite of some unusual results in the NPF groups for the A-AB group. In general, although there are some isolated statistically-significant results, they form no substantial pattern and so we conclude that these results support overall our previous conclusion that video game play has a negligible effect on performance in the TDT (Figures 2.2 and 2.3). See Table 2.4 for a full list of these results.

Main Effect of Video Game Play on:

	Baseline Day 1-A Performance		partial-η^2			
All Subjects	$F_{2,994}$ =	0.21	0.0003			
PF-Learning	$F_{2,581}$ =	2.27	0.0053			
PF-Transfer	$F_{2,525}$ =	2.52†	0.0060			
NPF-Learning	$F_{2,329}$ =	5.50**	0.0249			
NPF-Transfer	$F_{2,385}$ =	2.54†	0.0099			
	Learning : A-AB Subjects		**partial-η^2**	**Learning : AB-AB Subjects**		**partial-η^2**
All Subjects	$F_{1,308}$ =	5.16*	0.0135	$F_{1,343}$ =	0.02	> 0.0000
PF-Learning	$F_{1,168}$ =	0.72	0.0032	$F_{1,224}$ =	2.66	0.0095
PF-Transfer	[1] $F_{1,133}$ =	2.90†	0.0145	$F_{1,203}$ =	0.69	0.0028
NPF-Learning	[1] $F_{1,98}$ =	15.70***	0.0842	$F_{1,77}$ =	2.59	0.0198
NPF-Transfer	[1,2] $F_{1,133}$ =	4.28*	0.0209	[3] $F_{1,98}$ =	0.71	0.0042
	Transfer : A-AB Subjects		**partial-η^2**	**Transfer : AB-AB Subjects**		**partial-η^2**
All Subjects	$F_{1,308}$ =	0.33	0.0009	$F_{1,343}$ =	0.05	0.0001
PF-Learning	$F_{1,168}$ =	3.72	0.0165	$F_{1,224}$ =	1.39	0.0052
PF-Transfer	$F_{1,133}$ =	0.72	0.0034	$F_{1,203}$ =	2.28	0.0091
NPF-Learning	[1] $F_{1,98}$ =	10.89**	0.0720	$F_{1,77}$ =	1.79	0.0160
NPF-Transfer	$F_{1,133}$ =	0.08	0.0004	[3] $F_{1,98}$ =	2.30	0.0155

Table 2.4. Main effect of video game experience on baseline performance during Day 1-A session, learning, and transfer (split by experimental condition) for whole sample, PF, and NPF groups.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; † $p < 0.10$
[1] These analyses indicated a significant Experiment x Gaming interaction.
[2] This analysis indicated a significant Experiment x SOA x Gaming interaction.
[3] These analyses indicated a significant SOA x Gaming interaction.

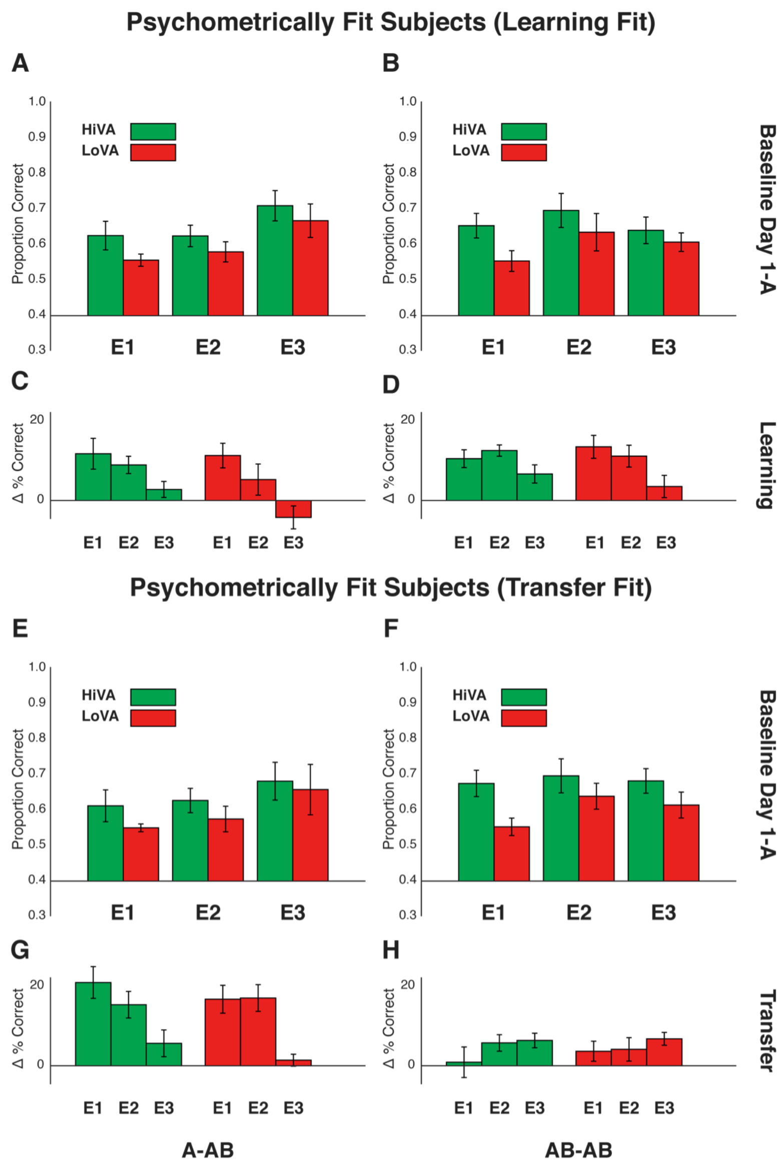


Figure 2.1. Baseline proportion correct (A,B) for PF-Learning subjects. Change in percent correct over these sessions (C,D). Baseline proportion correct (E,F) for PF-Transfer. Change in percent correct over these sessions (G,H). Subjects are divided by their visual attention skill (VA), with HiVA (green) and LoVA (red) indicated separately. Errorbars reflect standard error of the mean.

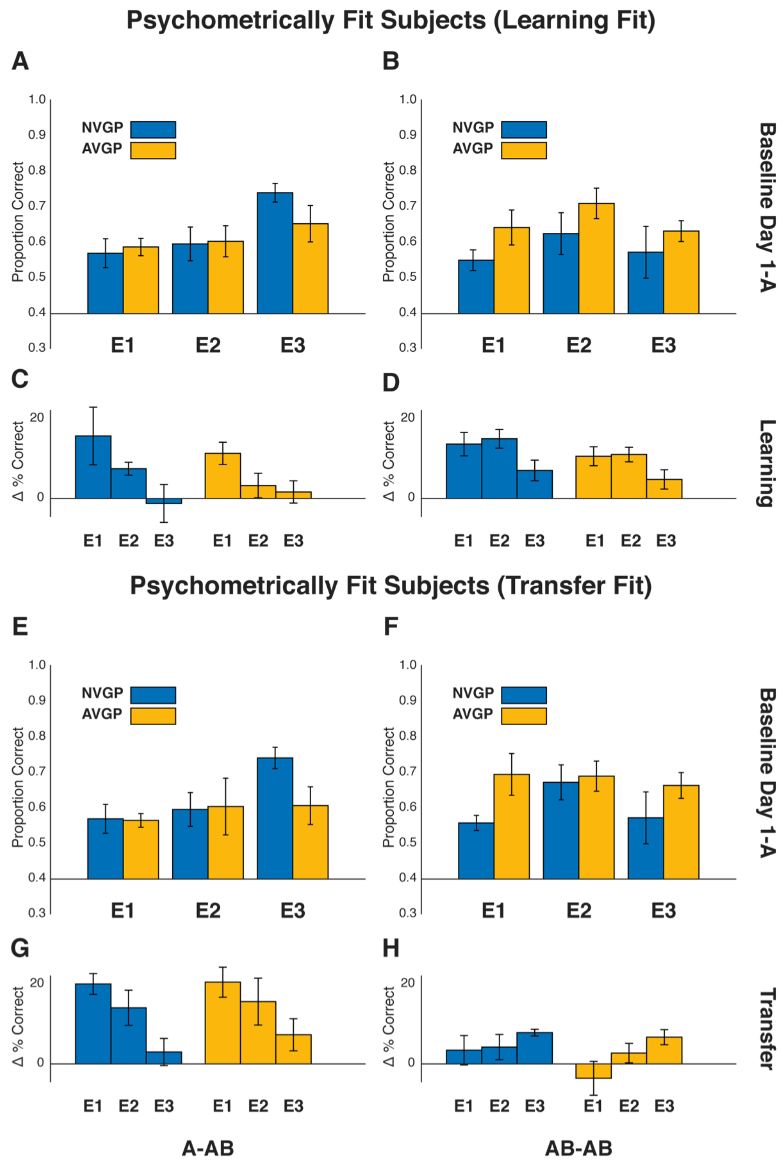


Figure 2.2. Baseline proportion correct (A,B) for PF-Learning subjects. Change in percent correct over these sessions (C,D). Baseline proportion correct (E,F) for PF-Transfer subjects. Change in percent correct over these sessions (G,H). Subjects are divided by video game experience, with NVGP (blue) and AVGP (yellow) indicated separately. Errorbars reflect standard error of the mean.

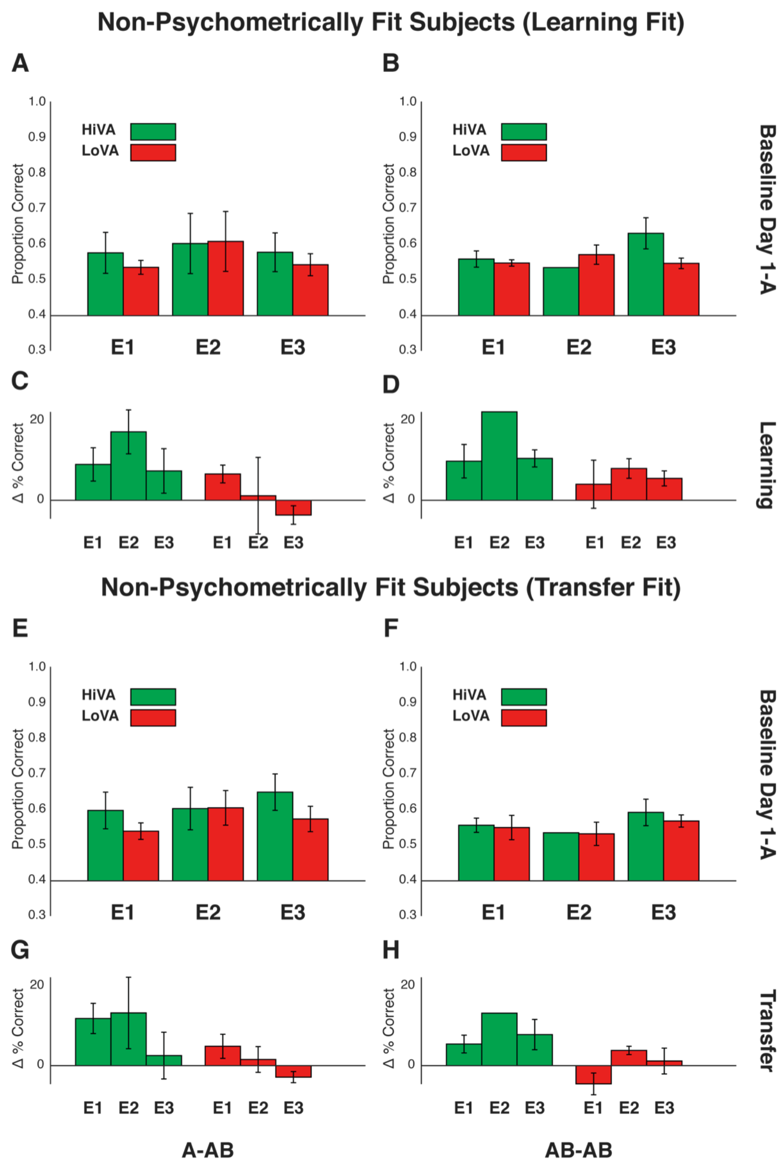


Figure 2.3. Baseline proportion correct (A,B) for NPF-Learning subjects. Change in percent correct over these sessions (C,D). Baseline proportion correct (E,F) for NPF-Transfer. Change in percent correct over these sessions (G,H). Subjects are divided by their visual attention skill (VA), with HiVA (green) and LoVA (red) indicated separately. Errorbars reflect standard error of the mean.

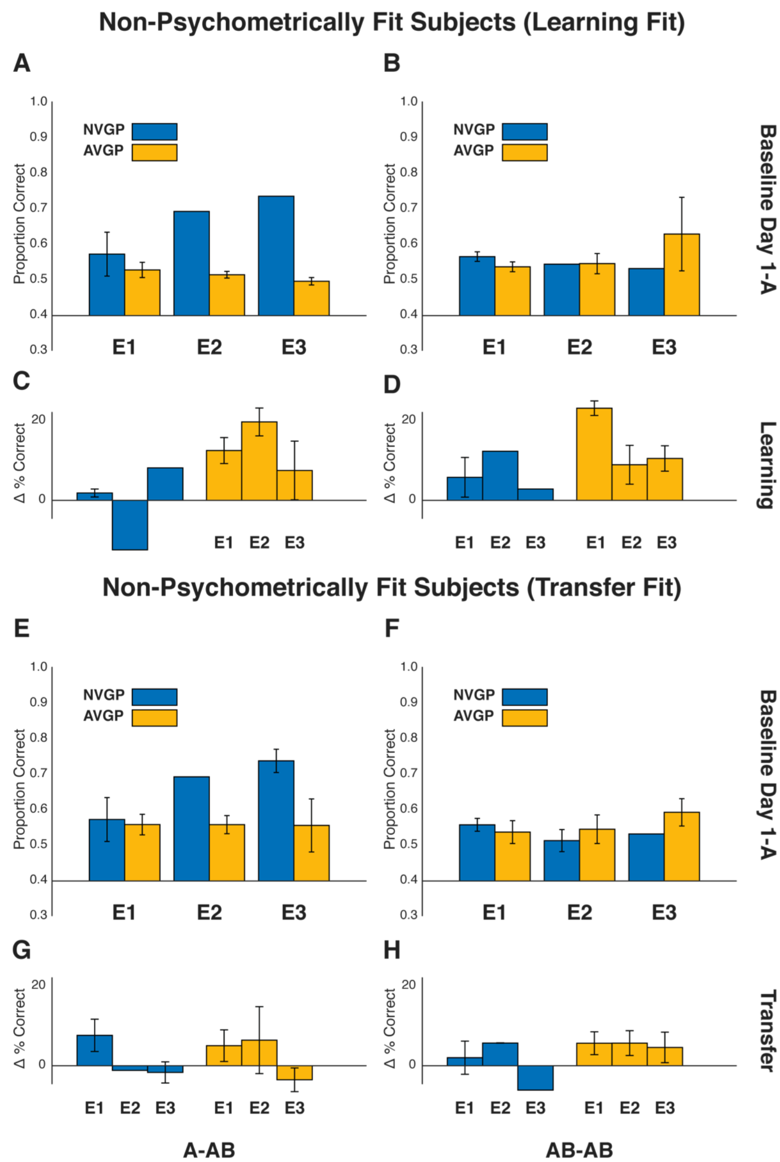


Figure 2.4. Baseline proportion correct (A,B) for NPF-Learning subjects. Change in percent correct over these sessions (C,D). Baseline proportion correct (E,F) for NPF-Transfer. Change in percent correct over these sessions (G,H). Subjects are divided by video game experience, with NVGP (blue) and AVGP (yellow) indicated separately. Errorbars reflect standard error of the mean.

Threshold Analysis

Any threshold-based analysis requires fitting in order to calculate the SOA value for a given performance threshold. As mentioned previously, many participants demonstrated particularly poor performance and may not have reached the necessary performance threshold of 70%. Alternatively, poor best-fits can result in nonsensical values for the calculated threshold without constraints on the fit. Therefore, we include the PF-Learning and PF-Transfer groups only for these analyses to ensure all subjects had a valid threshold for analysis. Results in the proportion-correct analysis refer to the results in Figures 2.5 and 2.6.

Moderating Effect of Roving the Order of Trial Difficulty

Baseline threshold results tell a somewhat mixed story relative to the proportion-correct based analysis. Although thresholds in Experiment 1 are lower (indicating better performance rather than worse, such as we as saw in the proportion-correct analysis), the PF selection process eliminated a large number of participants with diverse outcomes and preferentially retained subjects who may have performed unusually well on the more difficult Experiment 1 (Figures 2.5 and 2.6; A, B, E and F). Therefore, these results in particular should be viewed with skepticism. See Table 2.5 for a full list of these results.

	Main Effect of Experiment		partial-η^2	Experiment 1	Experiment 2	Experiment 3
PF-Learning	$F_{1,39}$ =	1.66	0.0326	98.5(25.2)	116.1(25.6)	116.1(25.2)
PF-Transfer	$F_{1,32}$ =	0.00	> 0.0000	98.7(31.7)	112.8(21.9)	116.1(25.2)

Table 2.5. Main effect of experiment on baseline performance on Day 1-A session with mean threshold (with SD) for each experiment for PF groups.

*** p < 0.001; ** p < 0.01; * p < 0.05; † p < 0.10

Moderating Effects of Visual Attention

The main effect of visual attention on baseline Day 1-A performance we observed in the whole-subject analysis for proportion correct is statistically significant in both PF groups when we look at the thresholds. We did not find significant effects of VA on threshold learning or transfer, with the one unexpected exception of a significant effect on transfer in one of the PF-Learning group analyses (Figure 2.5). These findings highlight our reasoning for reporting the proportion-correct based statistics in Chapter 1, the effects of visual attention on learning which are clear in the proportion-correct analysis are absent when we restrict ourselves to only a small subset of the overall dataset. See Table 2.6 for a full list of these results.

Main Effect of Visual Attention on:

	Baseline Day 1-A Performance	partial-η^2		
PF-Learning	$F_{1,39} =$ 5.20*	0.1022		
PF-Transfer	[1] $F_{1,32} =$ 9.54**	0.1864		
	Learning : A-AB Subjects	partial-η^2	Learning : AB-AB Subjects	partial-η^2
PF-Learning	$F_{1,15} =$ 0.19	0.0108	$F_{1,20} =$ 0.00	0.0001
PF-Transfer	$F_{1,10} =$ 0.82	0.0529	$F_{1,18} =$ 0.08	0.0037
	Transfer : A-AB Subjects	partial-η^2	Transfer : AB-AB Subjects	partial-η^2
PF-Learning	$F_{1,15} =$ 4.79*	0.1871	$F_{1,20} =$ 0.01	0.0005
PF-Transfer	$F_{1,10} =$ 0.93	0.0345	$F_{1,18} =$ 0.79	0.0353

Table 2.6. Main effect of visual attention on baseline performance during Day 1-A session, learning, and transfer (split by experimental condition) for whole sample, PF, and NPF groups.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; † $p < 0.10$

[1] This analysis indicated significant Condition x VA ($p = 0.036$) and Experiment x VA ($p = 0.015$) interactions.

Moderating Effects of Action Video Game Experience

The lack of a main effect of video game experience on baseline Day 1-A performance replicated in both PF groups in the threshold-based analysis (Figure 2.6). Unfortunately, our limited number of subjects with valid thresholds in the PF groups, when further divided by video game experience, makes some comparisons impossible due to invalid degrees of freedom, but the overall finding of no effects on learning or transfer from the proportion-correct analysis remains consistent. In general, these results support our conclusion that video game play has a negligible effect on performance in this task. See Table 2.7 for a full list of these results

Main Effect of Video Game Play on:

	Baseline Day 1-A Performance	partial-η^2		
PF-Learning	$F_{2,35} =$ 0.71	0.0312		
PF-Transfer	$F_{1,29} =$ 0.02	0.0005		
	Learning : A-AB Subjects	partial-η^2	Learning : AB-AB Subjects	partial-η^2
PF-Learning	$F_{1,8} =$ 0.45	0.0024	$F_{1,15} =$ 0.89	0.0398
PF-Transfer	$F_{0,3} =$ n/a	n/a	$F_{1,12} =$ 0.50	0.0362
	Transfer : A-AB Subjects	partial-η^2	Transfer : AB-AB Subjects	partial-η^2
PF-Learning	$F_{1,8} =$ 1.83	0.1634	$F_{1,15} =$ 3.12†	0.1292
PF-Transfer	$F_{0,3} =$ n/a	n/a	$F_{1,12} =$ 0.77	0.0502

Table 2.7. Main effect of video game experience on baseline performance during Day 1-A session, learning, and transfer (split by experimental condition) for whole sample, PF, and NPF groups.

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$; † $p < 0.10$

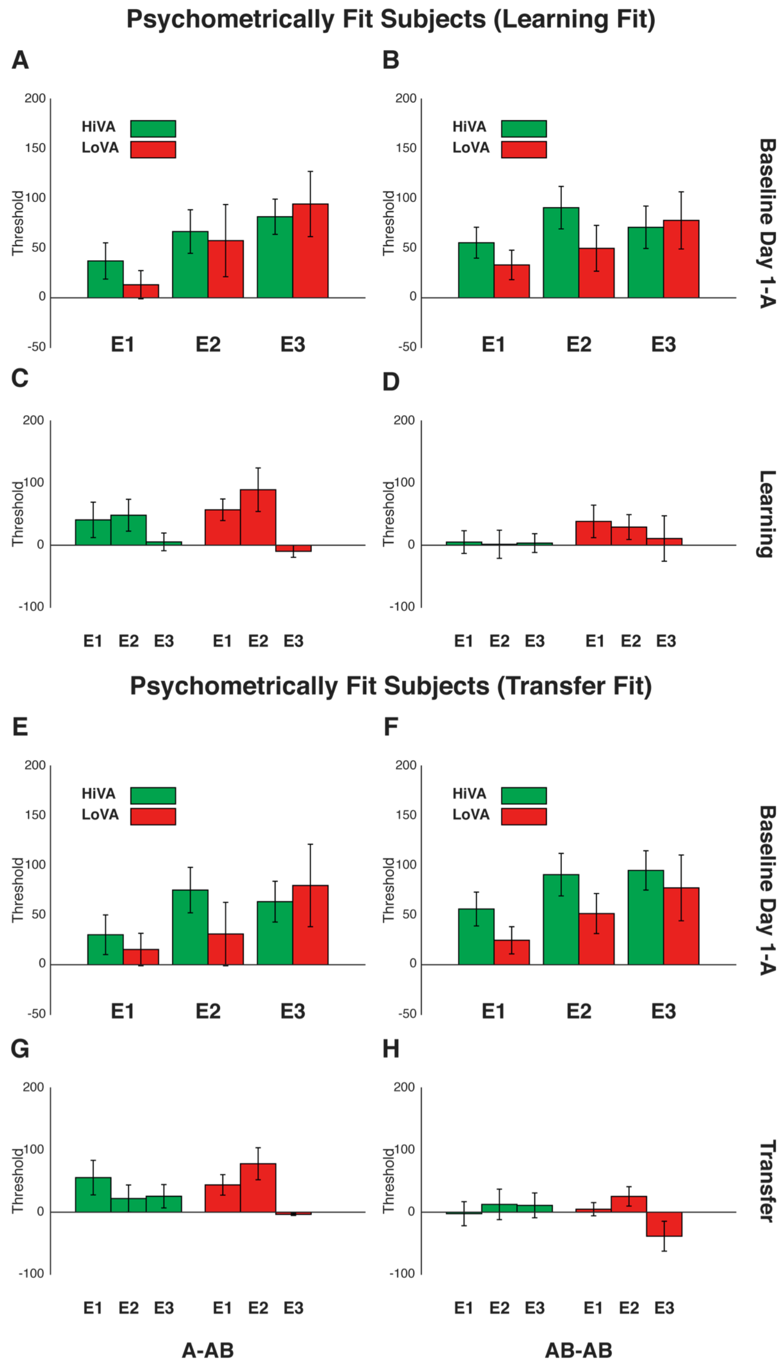


Figure 2.5. Estimated baseline threshold (A,B) for PF-Learning subjects. Change in estimated threshold between appropriate sessions for learning (C,D). Estimated baseline threshold (E,F) for PF-Transfer subjects. Change in estimated threshold between appropriate sessions for transfer (G,H). Subjects are divided by their visual attention skill (VA), with HiVA (green) and LoVA (red) indicated separately. Errorbars reflect standard error of the mean.

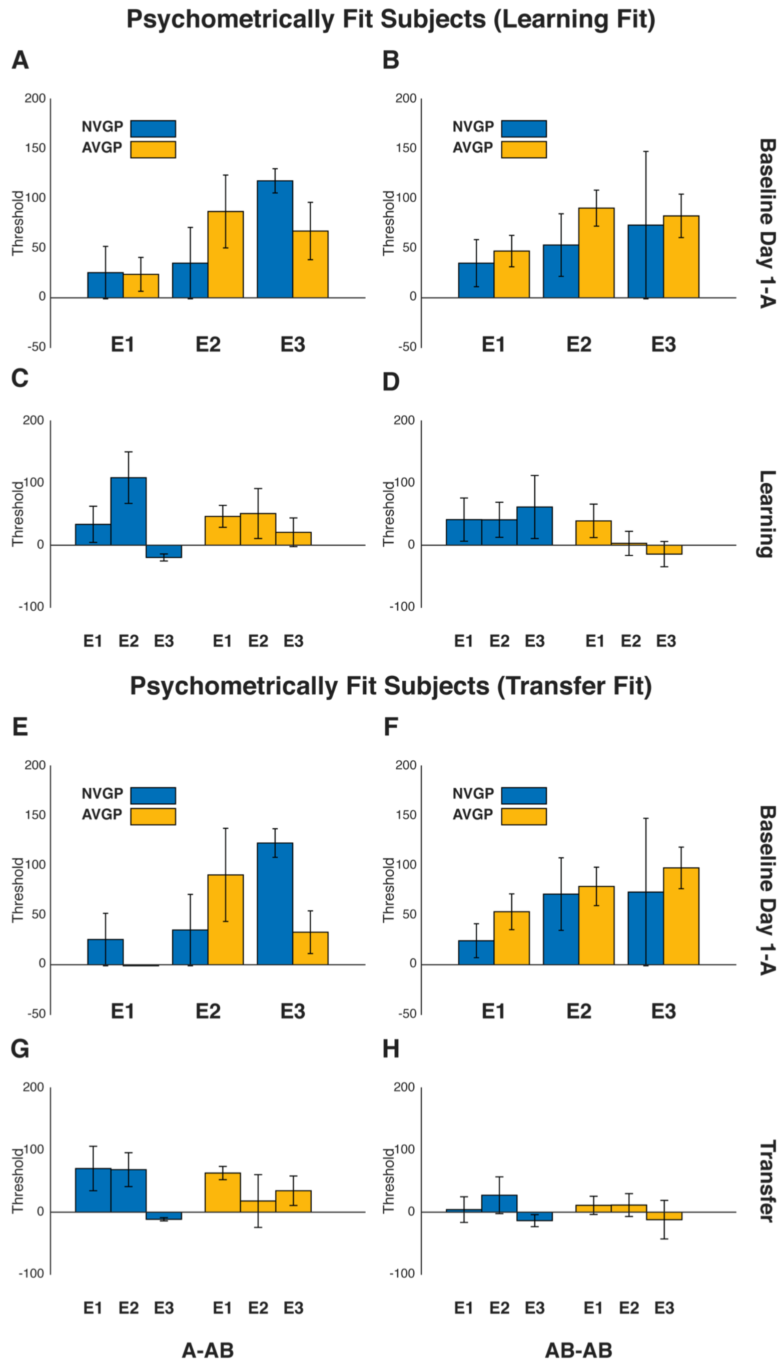


Figure 2.6. Estimated baseline threshold (A,B) for PF-Learning subjects. Change in estimated threshold between appropriate sessions for learning (C,D). Estimated baseline threshold (E,F) for PF-Transfer subjects. Change in estimated threshold between appropriate sessions for transfer (G,H). Subjects are divided by video game experience, with NVGP (blue) and AVGP (yellow) indicated separately. Errorbars reflect standard error of the mean.

Discussion

The purpose of the additional analyses conducted in this chapter was to explore the impact of inattentive and unfocused participants on the results of the large perceptual learning experiment described in Chapter 1. The impact of these additional analyses on our primary conclusions for the perceptual learning experiment discussed in Chapter 1 was minimal. With few exceptions the results from the PF/NPF analyses were consistent with those already reported based on the entire sample population. However, during the course of our follow-up analysis we faced the problem that the standard method of filtering out poorly-performing subjects was not a valid option due to the experimental manipulations of our study. As a result, we were obliged to develop novel analytical approaches to resolve this unique problem.

The first concern we addressed involved the definition of “bad” subjects who were inattentive and unfocused. There is some research studying the phenomenology of attentional lapses (Smallwood et al., 2004; Smallwood, 2011, Chernyshev et al., 2015) in general, but this research is not focused on the practical impact of these lapses on datasets in cognitive tasks. As we discussed before, there is some research investigating the numerical effects of these lapses (Harvey, 1986; Swanson & Birch, 1992), but authors such as Wichmann & Hill (2001) effectively take for granted that these lapses will occur. Omitting these “bad” subjects is the standard response to subjects with high lapse rates, but this approach can be inadequate. When an experimental manipulation induces reduced overall performance, such as roving the order of trial difficulty, the lapse rate becomes confounded with the experimental effects of interest.

Our solution to this concern was to re-define what it meant to be a “bad” subject. Mean and standard deviation are ubiquitous in psychology research as summary descriptive statistics, and conceptually they describe a one-dimensional average “expected” value and the variance of the date relative to that expectation. By using a constrained psychometric fit, we describe a two-dimensional expected *pattern* of behavior similar to a contrast test in an analysis of variance, and the MSE reflects the variance of the data relative to this pattern. Defining poor performance and “bad” subjects in terms of high-variance participants is a mathematically-sound alternative to the simple method of sorting out subjects with a low mean.

The second concern we addressed refers to the joint criterion necessary when we consider two sessions relative to one another in a dataset. The perceptual learning literature is full of examples of experiments where two sessions are compared to one another, but the need to selectively include participants on this basis is typically unnecessary. Simply replacing a subject with one poor session is easier than trying to salvage data that is potentially tainted by a high lapse rate. We were unable to follow this standard for the same reason we could not simply reject subjects based on low performance, and so developed an alternative approach. We are aware of the fact that the cutoff thresholds chosen do not perfectly match the proportions necessary to create a perfect median split. We are also aware the data does not satisfy assumptions of independence that would be necessary for this method to be formally valid. However, given the ultimate goal of merely dividing our subjects into two groups of approximately equal size we argue this approach passes the "sniff test" of face validity – our sample sizes following the application of this approach were indeed quite close to an equal split.

In the course of cognitive experimentation, it is important to conduct analyses that are both consistent with the existing standards in the field and conceptually valid for the types of data being examined. Most frequently, the standard method in the field is the most conceptually valid, and for good reason. However, when faced with a circumstance in which the standard in the field would not in fact be valid for a dataset, it is important to be flexible and to develop principled alternative data analytic approaches if necessary.

References

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

Chernyshev, B. V., Lazarev, I. E., Bryzgalov, D. V., & Novikov, N. A. (2015). Spontaneous attentional performance lapses during the auditory condensation task: An ERP study. *Psychology & Neuroscience*, *8*(1), 4.

D'Agostino, R. B. (1986). *Goodness-of-fit-techniques* (Vol. 68). CRC press.

Harvey, L. O. (1986). Efficient estimation of sensory thresholds. *Behavior Research Methods, Instruments, & Computers*, *18*(6), 623-632.

Karni, A, & Sagi, D. (1991). Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(11), 4966–70.

Smallwood, J. (2011). The footprints of a wandering mind: further examination of the time course of an attentional lapse. *Cognitive Neuroscience*, *2*(2), 91-97.

Smallwood, J., Davies, J. B., Heim, D., Finnigan, F., Sudberry, M., O'Connor, R., & Obonsawin, M. (2004). Subjective experience and the attentional lapse: Task engagement and disengagement during sustained attention. *Consciousness and Cognition*, *13*(4), 657-690.

Stickgold, R., James, L., & Hobson, J. A. (2000). Visual discrimination learning requires sleep after training. *Nature Neuroscience*, *3*(12), 1237.

Swanson, W. H., & Birch, E. E. (1992). Extracting thresholds from noisy psychophysical data. *Perception & Psychophysics*, *51*(5), 409-422.

Wichmann, F. A. & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293-1313.

Yotsumoto, Y., Chang, L. H., Watanabe, T., & Sasaki, Y. (2009). Interference and feature specificity in visual perceptual learning. *Vision Research*, *49*(21), 2611–2623.

Using Gaze Behaviors to Track

Changes in Perceptual Bias

During Perceptual Learning

Theodore Jacques[1] & Aaron R. Seitz[1]

Author Affiliations:

[1]Department of Psychology, University of California, Riverside, Riverside, CA 92521

Correspondence should be addressed to:

Theodore Jacques

theodore.jacques@gmail.com

(567)-203-8232

Abstract

We present a novel application of eye-tracking and image processing techniques to inform perceptual learning. This method exploits natural eye movements to examine changes in perceptual bias during visual search training. It uses fixations on a visual noise field to define regions of interest for frequency and orientation analysis, comparing relative power in non-target-fixation locations to the target frequency and orientation. In this paper we describe the method in detail, including an experiment conducted using this method. We evaluate the learning experiment with a variety of measures to illustrate the types of data our new method can collect to inform learning and suggest possible unsolved questions in the perceptual learning literature that may benefit from further investigation using this method. Finally, we discuss how our approach may be combined with traditional learning measures and other eye-tracking data to provide a unique perspective on the learning process that may be informative for a variety of perceptual learning questions.

Introduction

The field of perceptual learning utilizes a wide variety of experimental paradigms. This is a reflection of the broad scope of perceptual learning, which Gibson (1963) defines as "any relatively permanent and consistent change in the perception of a stimulus array, following practice or experience with this array (p. 29)." Gibson's characterization is perhaps the broadest possible way to interpret learning, since perception is a continuous process and humans are constantly adapting to our environment. The scientific study of perceptual learning encompasses every possible sensory modality, and every possible perceptual adaptation that one might observe in response to a changing environment. A complete review of the field of perceptual learning would be far beyond the scope of this work. However, we will discuss some of the most influential findings in the field in order to demonstrate what types of work have been done, the types of questions that the existing methods can address, and – importantly – illustrate the need for a new method to help answer questions which cannot be addressed with existing techniques.

Perceptual learning is predicated on the supposition that the brain can change in response to stimuli, and so some of the earliest roots of perceptual learning can be traced to Wiesel & Hubel's (1965) experiments on neural plasticity in kittens. Most of these experiments would not fit the modern definition of the perceptual learning training study, since they frequently interfered with the animals' original perceptions of the world from before the animals' eyes would naturally open. However, the basic observation that the brain is indeed plastic and changeable is one of the foundational assumptions of perceptual learning. Therefore, perceptual learning researchers began to explore different aspects of this plasticity. Is there a temporal window for the brain after which plasticity halts or is reduced (Flege, 1987)? To what extent is the learning we observe specific to different aspects of the stimulus or can it generalize to new stimuli, and what are the rules that govern specificity and generalization (Ahissar & Hochstein, 1997; Censor, Sagi & Cohen, 2012)? The observation that we can induce perceptual learning even in undergraduate students shows that perceptual learning can occur into adulthood. And although early findings suggested that perceptual learning may be highly specific to the particular aspect of the stimulus trained (Fiorentini & Berardi, 1980; Fahle, Edelman, & Poggio, 1995; Ball & Sekuler, 1987), later work has shown that this is not necessarily the case (Jeter et al., 2009; Xiao et al., 2008, Zhang et al, 2010). One over-arching conclusion that can be drawn from the thousands of papers published on perceptual learning is this: perceptual learning is tricky. The rules governing when we see specificity and when we see transfer are often highly specific to the particular task and stimulus being employed. Attention appears to play an important role in whether learning or transfer takes place (Sasaki, Nanez, & Watanabe, 2010; Wang et. al 2012) and so recent research is further investigating this relationship (see Donovan, Szpiro, & Carrasco, 2015; Galliussi et al, 2018; and Seitz & Watanabe, 2005 for a review).

However, in spite of the high degree of variability in the outcomes of perceptual learning studies, there are significant commonalities in the methods employed. Modern perceptual learning tasks predominantly rely on perceptual training paradigms in which subjects view a particular stimulus many times in a session and return for training on multiple consecutive days (Fiorentini & Berardi, 1980; Karni & Sagi, 1991; Weiss, Edelman, & Fahle, 1993). Researchers either plot performance from each session individually or make use of multiple training sessions to strengthen the learning effects they are looking for with a pre-test/post-test design. This convention is largely respected when the purpose of the study is to investigate the time course of learning (see Karni & Sagi, 1993; Tremblay, Kraus, McGee, 1998; and Atienza, Cantero & Dominguez-Martin, 2002 for a few examples). Even when we see truly out-of-the-box paradigms like Shibata et al.'s (2011) stimulus-free neurofeedback approach, incorporating neuroimaging techniques does not by itself mean that researchers have broken free of discrete testing sessions (Lewis et al., 2009). Although some recently published research incorporating some very exciting methods does deserve mention here. Kattner et al. (2017) utilize some novel approaches to estimate trial-by-trial performance by averaging across all their subjects; and using simulated observers rather than necessarily running human subjects can also provide insights into perceptual learning (Zhang et al., 2019).

One of the many topics of interest to researchers is the time course of perceptual learning. Within this category there are two types of learning over time considered. Some research focuses on the durability of learning and how long changes due to perceptual learning can endure (Yotsumoto, Watanabe & Sasaki, 2008; Qu, Song & Ding, 2010). Other research is more interested in the trajectory of learning over time as it is happening (Watson, 1980; Fahle, Edelman, & Poggio, 1995; Sireteanu & Rettenbach, 1995; Tremblay, Kraus, McGee, 1998; Atienza, Cantero & Dominguez-Martin, 2002). It is this second type of learning-over-time we are interested in addressing here.

Consider the following thought experiment: In order to measure whether or not learning has occurred we typically use at least two sessions: a pre-test conducted before perceptual training and a post-test. Perceptual training takes place over multiple training sessions spread across several days, and we measure performance for the trained stimulus or for a transfer stimulus before and after training. To measure the time course of learning, the obvious next step is to include an additional testing session in the middle of training. An additional data point mid-way helps plot the trajectory of learning but only roughly; more measurements would be required to infer additional details about what is going on during the learning process itself. But what would happen if as more and more testing sessions were added? For an experiment with training over four days, one additional test after day two seems reasonable, but a theoretical problem arises as we consider adding more and more tests. Perceptual learning training assumes that the changes we see in perception are a result of the training. Critically, however, task-related learning is also occurring as the subject is exposed to the testing task; this testing-related learning can account for a great deal of observed learning (Zhang et al., 2010). As we add more testing sessions, subjects will likely cross an invisible line where their improvement on the testing measure will be due more to experience with the measure itself than due to the training regime.

This presents a methodological problem. In order to understand the time course of learning we want to observe learning at many points as it is happening. However, if we stop the participant to measure their learning we run the risk of interfering with the process. The most common way to resolve this problem is with blocked measurements.

In his study of auditory perceptual learning, Watson (1980) computed the signal detection theoretic metric d’ (Green & Swets, 1966; Macmillan & Creelman, 2004) for each training day to calculate a psychometric function for learning. Karni & Sagi (1993) also used days as blocks in measures with the Texture Discrimination Task, and Tremblay, Kraus, McGee (1998) used a similar data-analytic approach, calculating average ERPs on a daily basis. In their study of vernier acuity, Fahle, Edelman, & Poggio, (1995) broke up their data into blocks of 80 trials and calculated a percentage correct for each block number. Effectively this approach treats individual blocks like miniature “days” of an experiment to look at performance over time. Atienza, Cantero & Dominguez-Martin (2002) did the same, looking at portions of a testing session while using EEG methods. It is not feasible, however, to simply define smaller and smaller block sizes, eventually “stimulus-independent lapses” (a polite way to talk about moments where the subject is not paying attention, Wichmann & Hill, 2001) will add too much variability to the measure and result in data too noisy to use. Fitting a function is an attractive choice, however this approach requires datapoints collected at high temporal resolution in order to make inferences about learning over short timescales. Aggregating information across subjects to make inferences about individual trials is a viable option (Kattner et al., 2017) but depends on strong assumptions that subjects will perform the task in similar ways. Continuous evaluation of an individual subject’s performance (as discussed by Seitz, 2017) is a more desirable option.

In this paper we will discuss a methodological resolution to this problem. In order to better characterize the short-term time course of learning within a session we need the highest possible temporal resolution while minimizing the impact of noise. We set out to design a novel paradigm that would fulfill the following requirements. First, it should allow for the highest possible temporal resolution for the data. Second, the approach should allow us to measure learning *in situ* without interrupting it for discrete testing sessions. Third, it should be a behavioral measure that depends on observable behaviors, rather than on inferences about brain activity. Fourth, it should not substantially impose on subject's natural learning processes and decision-making. And finally in addition to these aims, it should provide robust objective measures and rely on established data-analytic techniques to provide interpretable information.

To achieve these goals, we were particularly inspired by the work of Pärnamets et al. (2015), who exploited subjects' natural eye gaze behaviors to make inferences about moral decision making. Eye tracking as a tool of psychophysics already incorporates many of the requirements we set out to achieve. It is passive and non-interfering and also provides data at a very high temporal resolution. Eye tracking during the search process provides a wealth of data including eye position, velocity, fixation length and location, trajectory of eye gaze, and pupil data. Therefore, we developed a perceptual learning technique that capitalizes on the rich information contained in eye tracking data while still incorporating the task elements necessary to induce perceptual learning seen in standard behavioral training.

Our new method relies on a combination of eye tracking and image processing techniques in the context of a free visual search task. Subjects are presented with a visual noise field and instructed to locate and fixate upon a target embedded in the noise. See

Figure 3.1 for an illustration of the stimulus and how subjects will search for the target. For a concrete measure of learning we focus on the content of the background noise

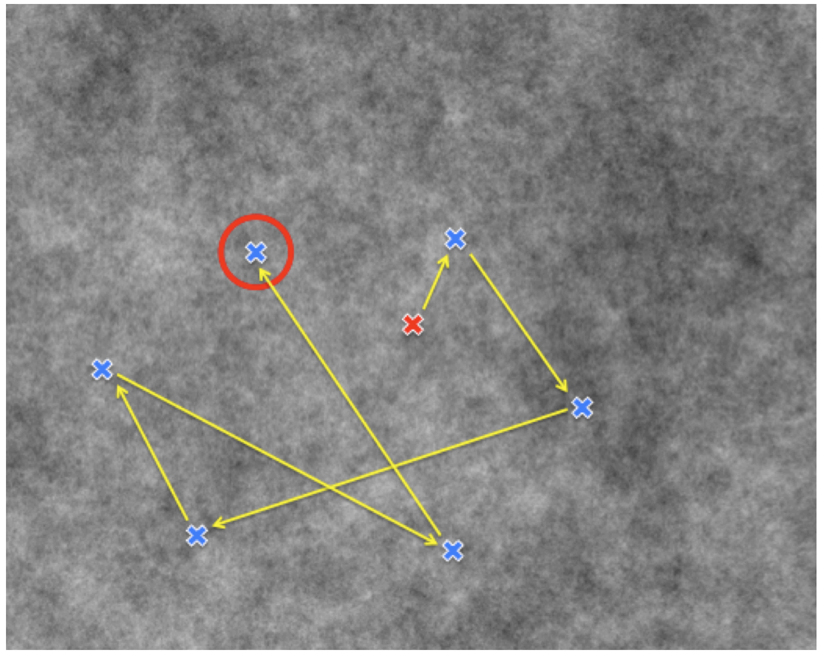


Figure 3.1. Example of a subject's gaze trajectory during the visual search task. Subjects begin in the center of the screen and freely make eye movements until the target is located.

at fixation points that are visited prior to locating the target. These non-target fixations define regions of interest, at which we apply Fourier image processing techniques to compare each region of interest to the target object and calculate a similarity score (See Figure 3.2). The visual search task is kept at a high level of difficulty with a staircase procedure, which ensures that there are always multiple non-target-fixations to evaluate. This provides within-trial data about the degree to which subjects are naturally choosing to fixate on portions of the background noise that resemble the target.

The ability to calculate within-trial measures is the key to improving temporal resolution in measuring perceptual learning. As we discussed before, attentional lapses can cause high single-trial variability in perceptual learning data. However, with our new method, failure to attend results in subjects simply failing to locate the target (and so it is easy to filter out lapse trials) and multiple fixations (and therefore similarity scores) per trial reduces variability in the averaged single-trial estimate of behavior. We can now theoretically reduce the block size down to a single trial lasting mere seconds and retain some confidence in the estimate of similarity score. Further, by adjusting the content of the target and the qualities of the background noise field it is possible to employ this approach to investigate a variety of aspects of perception.

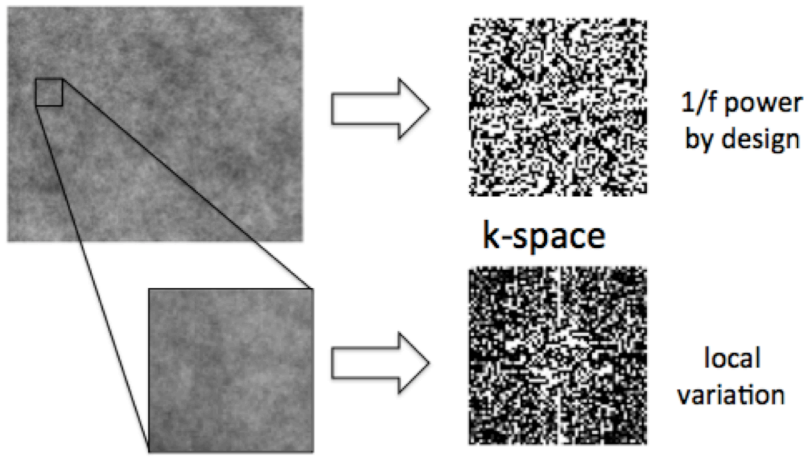


Figure 3.2. Illustration of patch analysis technique, the background image as a whole is 1/f by design but selected regions show local variation and are compared with the target object to compute similarity scores.

To our knowledge we are the first to take advantage of non-target fixations in this way to generate classification images for use in a perceptual learning context and the first to develop a protocol robust enough to provide consistent results and generalizable enough for application to a variety of research questions. We first introduced this technique in a presentation at the 2018 annual meeting of the Vision Sciences Society (Jacques & Seitz, 2018). Here we expand on the method in detail sufficient for others to apply to their own research. In addition, this paper discusses other methodological problems we resolved in pursuit of developing this approach. Finally, we include results from our own experimental program using this method in a perceptual learning context and provide data illustrating some of the ways to utilize this approach.

Methods

Visual Search Task Overview

Subjects are instructed to locate a target (gabor patch) embedded at a random location in a visual noise field. At the beginning of each trial, subjects view a circular focus in the center of a uniform gray screen and fixate upon it. We use the eye tracker to ensure central fixation before the trial begins. Once the noise field appears, subjects are free to look anywhere on the screen and are given no instructions except to fixate on the target once they locate it. After the subject has located the target and fixated upon it for a sufficient period (approximately 33ms), the eye tracker ends the trial immediately and the subject receives positive auditory feedback. No keyboard or verbal response is necessary for the task. If the subject is unable to locate the target within twenty seconds the trial ends automatically and the subject receives negative auditory feedback. The target is embedded in the noise field using alpha blending, and the transparency of the target is a value between zero and one. To ensure that subjects are always actively searching for the target (and producing multiple non-target fixations) we use a staircase method to control the difficulty of detecting the target by adjusting the transparency dynamically. To facilitate perceptual learning subjects always search for a target gabor of the same orientation and spatial frequency, and they train in this way for approximately one hour per day for 5 days (approximately 270 trials per day).

Stimulus Description

Target

At least one target is required in order to complete the visual search task. For our initial validation of the method we chose to use a Gabor patch (sinusoidal grating) as the target. Gabor patches are frequently used in perceptual learning research, and contain spatial frequency and orientation information similar to the 1/f noise field. Gabor patches are useful in this application because in k-space a Gabor patch contains significant energy at a particular spatial frequency and orientation, much more so than a simple line. This strong signature facilitates calculating the similarity score between the target and non-target locations selected by subjects during training. Alternative research questions may require different target stimuli, but a Gabor is an ideal target for basic perceptual learning applications. The target appears in a random location on the screen with the exception of two exclusion zones for the stimulus display which prevent the target from appearing either in the exact center of the screen (where subjects would not need to make an eye movement to be fixating upon it) or so close to the edge of the screen that it would be partially off-screen. Subjects may be informed of these restrictions; to our knowledge this information does not impact their behavior during the task.

Noise

We create noise fields using a random number generator set to produce an array of random grayscale luminance values of a uniform distribution. We treat the noise image as being generated in k-space, and apply a filter to create the desired noise type. The type of filter applied would vary based on the theoretical needs of the research question, but for most applications we recommend a 1/f filter. This filter is appropriate for use with the Gabor patches we selected for the target object in our experiment, alternative target objects with different visual properties would likely require noise backgrounds with different statistical properties and alternative filters. The final step is to apply an inverse Fourier transform to convert the k-space noise into an image for display.

The content of the background image is critical for data analysis. Therefore, the noise images themselves or the random seed used to generate the noise should be saved for later data analysis. In our experiment we found that due to hardware limitations it was helpful to create a library of noise images in advance of stimulus presentation and record the identity of the background image used for each trial. This was necessary because the inverse Fourier transform to create the final image can be computationally demanding and the size of the image scales with the square of the chosen display's screen resolution. Thus the inter-trial-interval necessary to accommodate these calculations was longer than desired while loading pre-generated noise fields is faster. We stored our noise images in batches of twenty trials, and during the task the inter-trial-interval was extended after every twenty trials while the next batch of noise images were loaded from memory. Subjects were instructed to briefly rest their minds without moving from the eye-tracking apparatus, and these pauses typically lasted only 5-10 seconds. The requirement to fixate on a circular focus before each trial begins ensured that subjects are attending to the task before the experiment proceeds.

Image Composition

We construct the visual search stimulus using alpha blending. Alpha blending is a common image processing technique for combining two images to produce a "transparency" effect, although it is used infrequently in perception research. For example in the field of visual contrast there are several methods of combining images at different contrasts, however most of these methods focus on blending one image into a uniform gray background and not a non-uniform background such as the noise field (see Wu & Tsai, 2003 for one such approach). This is unsuitable for combining a target with a non-uniform background like 1/f noise. In alpha blending, the final pixel luminance is the weighted mean of the luminance of the contributing images and the weight is referred to as the alpha level. For example, for a totally-black pixel blended with a totally-white pixel for an alpha of 0.4, the final pixel luminance would be 0.4x0 + 0.6x1, or 0.6. Therefore, while alpha blending is not frequently seen in visual perception research it is appropriate for this application where we are combining two complex images.

We create the final stimulus image by successive application of alpha blending to build the stimulus layer by layer. The base layer is a uniform gray background, and we layer first the noise field and then the target into the image. We define the alpha level for each blended layer independently, which provides the added benefit that these levels can be adaptively manipulated as needed. There is no limit to the number of objects that may be added to the image, even if they are overlap. For our validation measure we adaptively changed the alpha level of the target gabor while keeping the alpha for the background noise constant, but researchers could vary the alpha of the noise or other image elements as needed to address particular experimental questions.

Eye Tracking

General

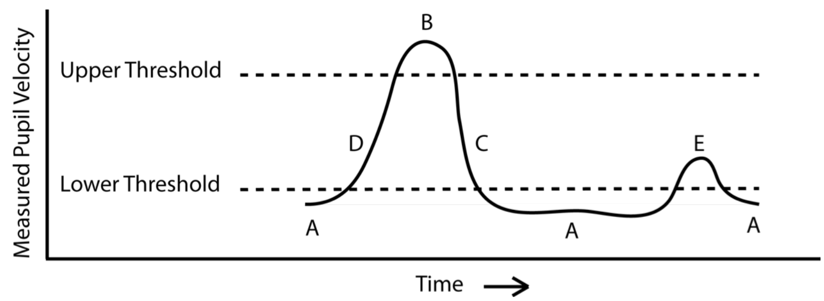
As part of the development of this overall technique, we also needed to complete a number of smaller projects. The most significant of these is the fixation detection algorithm. In eye tracking research, the most expedient way to identify if a subject is looking at a target is to define a region around the target and record the moment-to-moment eye position. If the eye position remains within the region for a critical period of time, the subject is assumed to be fixating at the desired location (see Lykins, Meana, & Kambe, 2006 for one such example using “Scene Regions”). This is convenient when researchers are interested in identifying when subjects are looking at a known location but is inadequate for identifying general fixations. There are a variety of algorithms for detecting general fixations (see Nyström & Holmqvist, 2010 and Hessels et al., 2017 for specific examples and Salvucci & Goldberg, 2000 for a review of the topic) and many eye-tracking manufacturers include proprietary software for classifying eye movement behaviors including fixations. However, a common theme among these approaches is that they are applied in data post-processing after the subject has completed the task. Salvucci & Goldberg (2000) explain that the most common methods for determining fixations are temporal methods based on calculated pupil velocity. In order to reduce noise and smooth the velocity estimate most methods use a running average approach with a window that includes both past and future datapoints to calculate the instantaneous speed at each time point. While desirable for reliable fixation detection, the use of prospective data in the velocity estimate limits researchers to post-hoc analyses only. In our task we need to know information about fixations both at the target location and at unforeseen locations. And while we could use the "scene region" technique for locating the target and another method for general fixation detection at other locations, we were reluctant to use a mix-and-match approach. Therefore we developed an algorithm for general fixation detection that utilizes only retrospective information that we can use throughout the task. This approach allows us to identify when a fixation is occurring in real-time, and verifying the location of the fixation for determining if the target has been located is straightforward.

Fixation Detection Algorithm

Our algorithm is a type of I-VT method as described by Salvucci & Goldberg (2000), with the addition of some elements of dispersion-based (I-DT) methods to account for slow drifts in eye movement. We note that combining I-VT and I-DT methods is not frequently reported in the literature on eye tracking and gaze determination (Andersson et al., 2017), and the use of both approaches together is itself a novel contribution to this field (if a small one).

We begin by calculating the point-to-point velocities for each datapoint we collect, and compare it to two velocity thresholds. Due to the fact we only use retrospective methods this velocity estimate is necessarily noisy; however, since our objective is only to define behaviorally-significant regions of interest then some variance in this estimate is not prohibitive. See Figure 3.3 for a diagram of the logic used to determine the identity of each data point vis-à-vis its status a fixation, saccadic eye movement, or other gaze behavior. The first velocity threshold is comparatively low, and datapoints below this threshold are assumed to be part of a fixation, with one small caveat we will discuss later. The second velocity threshold is quite high, high enough that we can safely assume

A



B

1. IF : Velocity < Lower Threshold
   SET : Datapoint Label = Fixation — Points A
   SET : Most recent threshold crossed = Lower Threshold
2. IF : Velocity > Upper Threshold
   SET : Datapoint Label = Saccadic eye movement — Point B
   SET : Most recent threshold crossed = Upper Threshold

3. IF : Velocity > Lower Threshold AND Velocity < Upper Threshold — Points C, D, E
   a) IF Most recent threshold crossed = Upper Threshold
      SET : Datapoint Label = Saccadic eye movement — Point C
   b) IF Most recent threshold crossed = Lower Threshold
      WAIT until second threshold is crossed
      c) IF second threshold crossed = Upper Threshold
         SET : Datapoint Label = Saccadic Eye movement — Point D
      d) IF second threshold crossed = Lower Threshold
         SET : Datapoint Label = Micro-saccidic eye movement — Point E

Figure 3.3.
A. Diagram of a sample velocity trajectory during an eye-tracking task with velocity thresholds included. Letters indicate distinct portions of the trajectory over time.
B. Logical flow-chart to decide the fixation-status of each datapoint over time, note that only velocity of the present datapoint is required for the purposes of establishing the beginning or end of each fixation event and any ambiguity in gaze behaviors occurs only during saccadic and micro-saccadic events.

that only data points corresponding to gross eye movements like deliberate saccades will surpass this threshold. Datapoints falling between these two thresholds are identified based on the most recent threshold crossed, or we allow a brief period of ambiguity before retroactively tagging the datapoints. We resolve the identity of each of these ambiguous (between-threshold) points by checking which threshold was crossed most recently and/or by waiting for the next crossing. If the most recent threshold crossed was the higher threshold, we infer the velocity is dropping at the end of a saccadic eye movement and append the current datapoints to that saccade. In contrast, if the most recent threshold crossed was the lower threshold, we wait for the next threshold to be crossed. If the next threshold is the higher threshold then we infer that the increase in velocity marked the beginning of a saccade, while if the lower threshold is re-crossed then we infer a petite eye movement (i.e., microsaccade) occurred. We then retroactively tag the ambiguous datapoints with these labels. This algorithm is not ideal for strict identification of all types of gaze behavior, however since our objective is to identify fixations reliably some temporary ambiguity is not a problem.

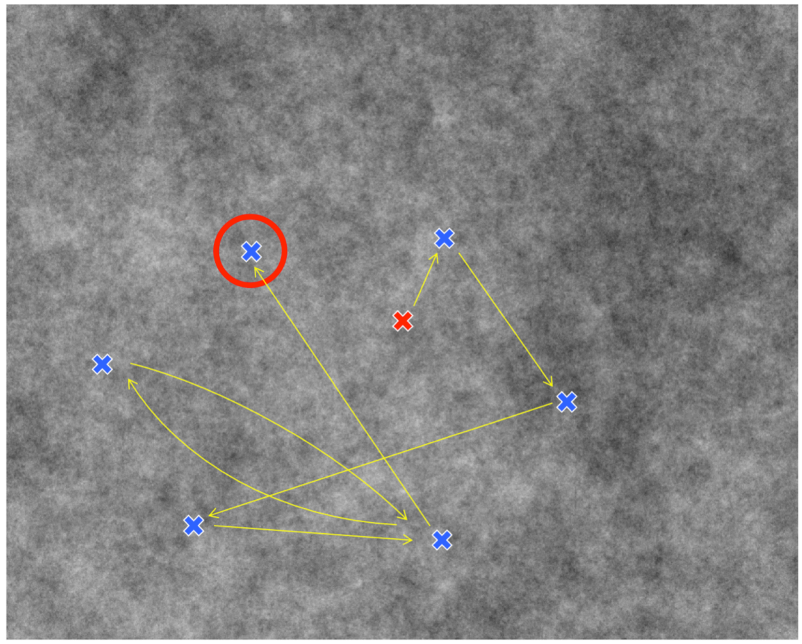
The caveat to the assumption that data points with low velocities are part of a fixation is the fact that we incorporate a location-sensitive measure in the fixation algorithm in addition to the velocity-based one. After the lower threshold has been crossed, we include a brief confirmation period of 33ms to ensure the fixation has truly begun. At a scanning rate of 100hz, this corresponds to about 3 consecutive datapoints below the lower velocity threshold. Based on those few datapoints we calculate a set of mean x and y reference coordinates for the fixation to provide a rough estimate of its location. After the fixation has been identified, each subsequent datapoint in the fixation is compared to the fixation location estimate, and if the eye has drifted more than 1 degree of visual angle from that point (even if a saccade event has not been recorded) we terminate the fixation and define the beginning of another fixation from that point. This accounts for slow eye drifts, but in practice we find these events to be quite rare in this task since slow drifting eye movements are not desirable for subjects to make in an active visual search task.

Eye Tracking Post-Processing

The fixation detection algorithm described above is useful for identifying fixations in real-time and suits the need to identify regions of interest for later data analysis. However, before we conduct further analysis we apply additional data processing steps to reduce the noise of the gaze data and provide further insight. Although we use a rough estimate of fixation location for the purposes of fixation detection, first we must calculate a more reliable centroid for the fixation during post-processing using all data points in the fixation. Second, the fixation detection algorithm does not account for previous fixation locations and due to the noisy velocity estimate sometimes long fixation events are broken into two or more parts due to microsaccadic movements (or perhaps mechanical noise in the data) even though the location of the "separate" fixations overlap. These "separate"-but-actually-continuous fixations are consolidated together and treated as one. Third, based on the unique locations of each fixation we calculate a trajectory for the eye movements and identify when two fixations that are not temporally adjacent actually refer to the same spatial location (See Figure

3.4). This accounts for situations where subjects return to the same point of interest for a second look. Fourth, we calculate the total dwell-time for each fixation event and also for each region of interest. This second measure accounts for the combination of dwell-time in the same region for separate visits. This additional post-processing provides us with information on the total number of unique non-target fixations, the order in which they were visited, and the length of time spent at each location. By itself, this data is valuable for understanding learning behavior, but it can also be incorporated in the image content analysis.

A



B

1. Identify location of every fixation recorded during the trial by taking the mean x and y coordinate of all datapoints associated with that trial.
2. If a fixation is in the same location as the previous fixation, combine the two and treat as one fixation.
3. Identify the trajectory of fixations by ordering them, and account for return visits to the same location when calculating dwell time.

Figure 3.4.
A. Example of a subject's gaze trajectory during the visual search task where the subject returns to a previously-visited location during the trial prior to locating the target.
B. Step-by-step process of organizing fixations within a trial for later data analysis.

Calculating Similarity Scores

After identifying non-target fixations and plotting the trajectory of eye movements, the next step in data analysis is to calculate a similarity score between each non-target fixation and the target. First we extract a square patch of the image surrounding the centroid of the fixation. Next, we use a Fourier transform to convert the image into k-space. This is necessary because we are interested in the degree to which the noise image contains spatial frequency and orientation information similar to the target, rather than whether or not the noise region superficially resembles the target. Finally, we compare the k-space noise image to the k-space version of the idealized target without noise, at this step we do need to account for phase differences between the target object and the non-target fixation. A gabor patch has a strong signature in k-space corresponding to its orientation and the fundamental and harmonic frequencies of the gabor's frequency. We calculate this similarity score using the inner-product or dot-product. The dot-product is calculated by multiplying two matrices together point-by-point and summing the total into a single value; this technique effectively identifies whether or not the two images share significant energy at the same points in k-space (see Hefferon, 2017 for a complete review of the topic). We calculate the similarity score for each non-target fixation in a trial.

We can combine similarity scores between non-target fixations within a trial to reduce the noise in our trial-by-trial estimate of behavior. If desired, we can also combine trials together into blocks. However this method differs from other block-based perceptual learning approaches due to the fact that the similarity score is not a measure of subject performance like accuracy or reaction time. The similarity score provides an estimate of both the subject's perceptual bias towards the stimulus being presented and any conscious strategies they may be employing to complete the task, and we can track changes in that aspects of behavior over time, but the similarity score does not correspond to any particular task-related behavior. Therefore, similarity scores are similar to ERP data, pupil data, or BOLD data in the way we interpret our results, in that they allow researchers to draw inferences about changes to mental processes. Even though the similarity score is not an estimate of observed behavior, it is derived entirely from observations of subject behavior and thus makes no assumptions about the inner mental processes of the subject or transformations between indirect measurements of neural activity (e.g., via EEG or fMRI) and the presumed underlying neural activity itself.

Experimental Task Methods

Subjects and Procedure

To validate the efficacy of the method to induce perpetual learning we used a traditional pre-test/post-test design to obtain measures of learning and transfer. The study took place over the course of seven consecutive days, where days 1 and 7 were testing days and days 2-6 were training days. Subjects were recruited using fliers posted on campus at the University of California, Riverside, and all subjects were UCR undergraduates. Some subjects (approximately 4) were excluded due to technical difficulties with the eye tracker that arose following recruitment. The most common type of technical difficulty was the eyetracker being unable to get a reliable fix on the pupil location. This could occur if the contrast between pupil and iris was inadequate to maintain a lock on the pupil or in one case a subject had exceptionally long eyelashes that caused the eye tracking software to mistake a clump of lashes for the pupil and the issue could not be resolved. Therefore, we report data for a small sample of 6 subjects. All subjects reported normal or corrected-to-normal vision and indicated they were able to see all test stimuli with no difficulty. Subjects provided informed consent prior to the beginning of the experiment and were compensated $10 per hour for their participation.

Apparatus

An Apple Mac Mini running Matlab (Mathworks, Natick, MA) and Psychtoolbox Version 3 (Brainard, 1997; Pelli, 1997) was used for stimulus generation and experiment control. Stimuli were displayed on a 16-inch Viewsonic PF817 monitor at a resolution of 1280x1024 pixels at 100 Hz by an NVIDIA GeForce 9400 graphics card (NVIDIA Corporation, Santa Clara, CA) and attenuated using a Bits++ system (Cambridge Research Systems, Cambridge MA). Eye tracking data was collected using a ViewPoint EyeTracker (Arrington Research, Scotsdale, AZ). Subjects were seated with their chins in a chinrest, and viewed the stimulus at a distance of 75cm in a darkened room. The eye tracker was set to allow inaccuracy up to 1.5 degrees of visual angle.

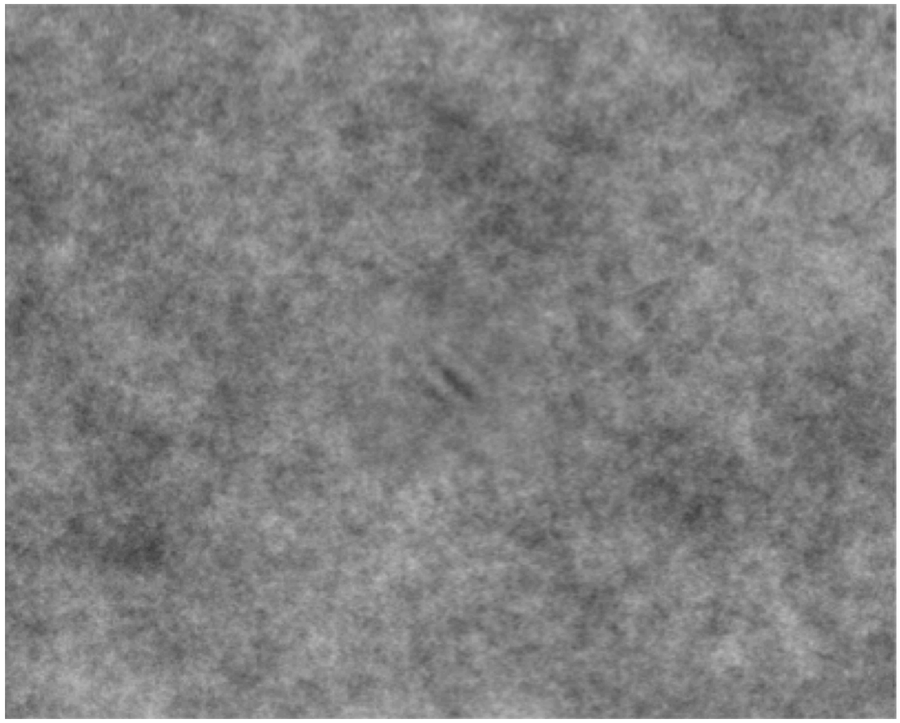
Stimulus

The target objects were Gabor patches at 0.5, 2, and 8 cycles/degree, rotated either 45° or 135°. Noise images consisted of 1/f random noise, and gabor patches were embedded in the background noise using alpha blending.

Pre-test and Post-test

Pre-test and post-test sessions were identical both in procedure and stimulus. The eye tracker was not used during these test sessions. Each trial began by presenting a circular focus in the center of the screen, and subjects were instructed to practice looking at the focus between trials because the eye tracker would enforce fixation on the training days. The focus was visible for 1 second before the trial began. Then subjects were presented with a random noise image drawn from the library of available noise image and the target Gabor was embedded in the noise at the center of the screen (see Figure 3.5 for an example stimulus and a diagram of the task.).

A



B

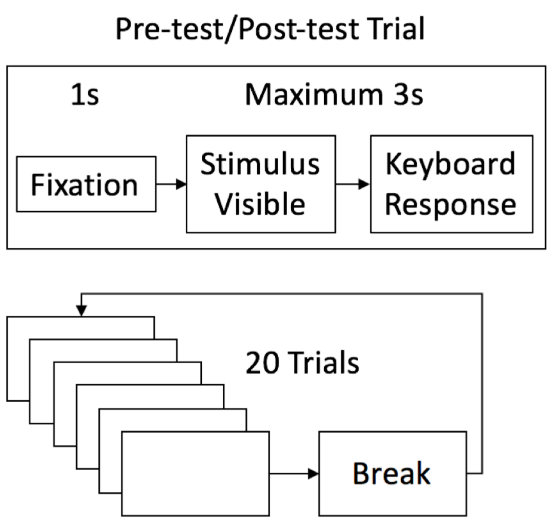


Figure 3.5.
A. Sample pre-test/post-test stimulus. The gabor is presented centrally in the noise field and subjects respond indicating the orientation of the gabor.
B. Outline of the order of the task.

Subjects were instructed that the target would always be presented centrally. Subjects responded indicating the orientation of the gabor using the keyboard. Subjects were allowed 3 seconds to make a response or the trial was automatically considered incorrect. Every 20 trials there was an extended break between trials while the next image library was loaded from memory, subjects were instructed to rest their minds briefly but remain vigilant to the task.

The alpha level of the target gabor varied according to 12 independent randomly interleaved 3-down, 2-up staircases (with streaking allowed), two for each spatial frequency/orientation combination. Each staircase started at an initial alpha level of 0.35, and varied by 0.2 per step for the first 4 reversals per staircase and then by 0.05 per step. The session ended after all 12 staircases had reached 8 reversals, or if each staircase had accumulated 60 trials. To determine the order of the interleaved we used mini-blocks of 12 trials presented in a random order, and repeated mini-blocks until the end of the session. This session usually took less than one hour. After the end of the session, the two staircases for each spatial frequency/orientation combination were used to calculate a single threshold value for that stimulus. Most often this was the mean of the two estimates, but in the event of unusual variance in one staircase the more estimate with lower variance was used. We recorded both the threshold itself and the standard deviation of the last 5 reversals; this information was used for the training sessions. Thresholds for the trained spatial frequency/orientation combination provide estimates of near-transfer learning (a centrally-located target is technically a transfer task relative to a randomly-located target in the visual search task) while untrained combinations provide estimates of far-transfer.

Training

Each training session followed the same procedure. The session began by calibrating the eye tracker and ensuring the subject was comfortable for an extended period. At the beginning of each session an example of the target gabor was shown to the subject at high contrast, to remind them of the target they needed to locate during the visual search task. Following this, each trial followed the same procedure. The fixation circle appeared in the center of the screen and subjects were instructed to fixate upon it. If subjects failed to fixate within 15 seconds, the experiment stopped and the eye tracker was re-calibrated. Otherwise, the trial began immediately following the detection of the fixation. The

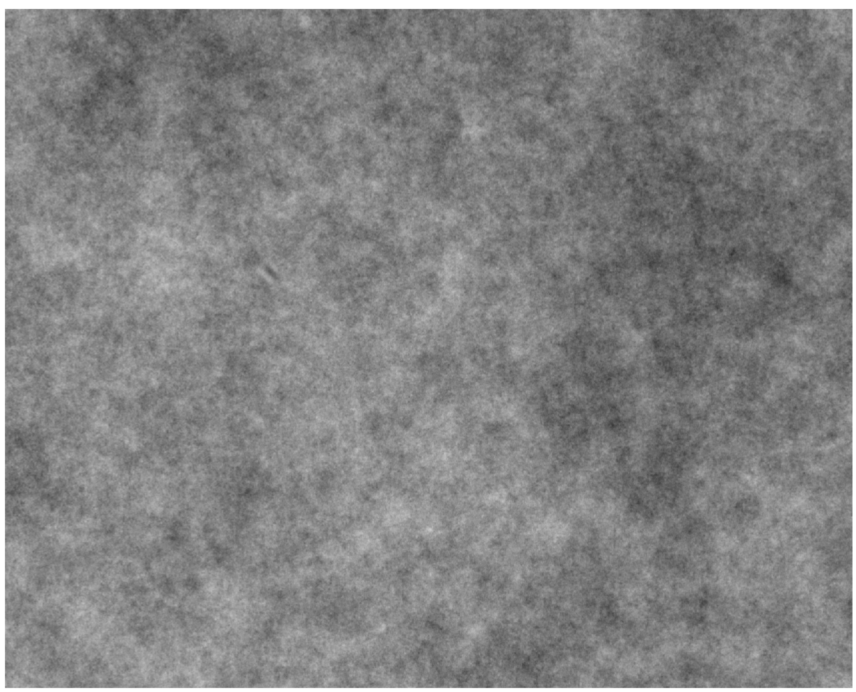


Figure 3.6. Sample stimulus from the visual search task, the subject must locate the gabor embedded a low alpha level at a random location in the noise field.

stimulus was presented with the target gabor embedded at a random location (See Figure 3.6 for an example stimulus). The initial alpha level of the target was set at 1 standard-deviation above the calculated threshold for that subject based on their threshold determined at pre-test, and was varied according to a 3-down, 2-up staircase with streaking allowed. The alpha level varied by 0.2 per step for the first 4 reversals and then by 0.05 per step for the remainder of the training session.

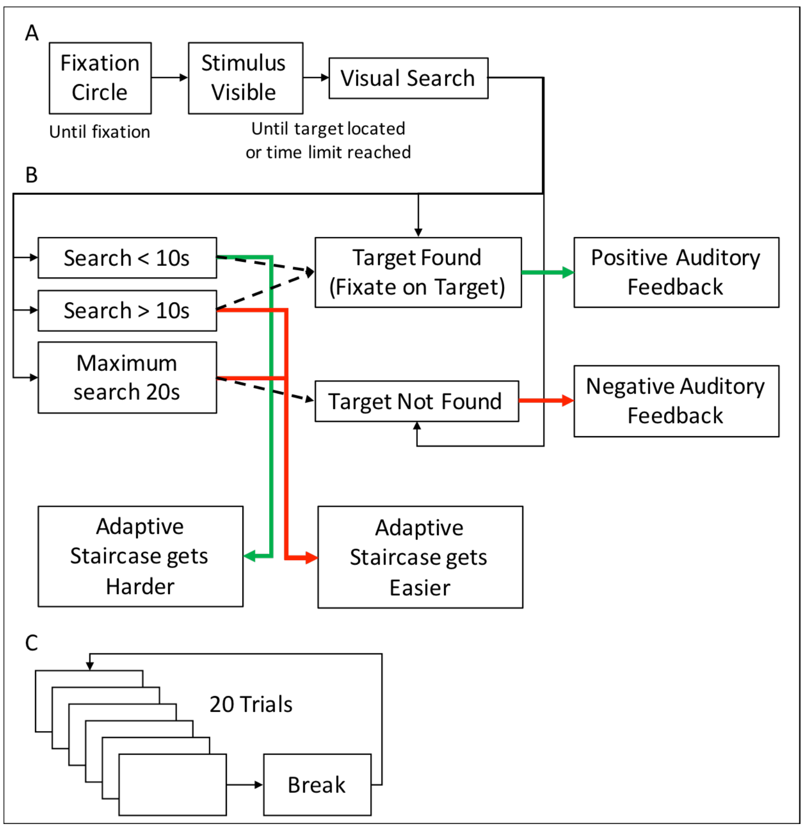


Figure 3.7.
A. Order of events in a single trial.
B. Order of trial outcomes, subjects receive positive auditory feedback any trial where they locate and fixate upon the target. For some trials this is false feedback, the adaptive staircase responds to the time required to locate the target.
C. Trials are presented in blocks of 20 with a brief break in to allow loading in of new background noise images for display. The process is repeated until the end of the session (approximately 270 trials/day).

Subjects were given 20 seconds to locate the target and fixate upon it. Subjects made no responses using the keyboard during training; the task is completed exclusively based on eye movements. We used false feedback to encourage subjects: if the target was not located within 20 seconds the subject received negative auditory feedback, whereas if the target was located within that time the subject received positive auditory feedback. For the staircasing procedure, only responses within 10 seconds were treated as correct to ensure the task never became too difficult. We calculated a threshold each day as a measure of task learning, and recorded overall accuracy, time to locate the target, and fixation-related data for each trial (see Figure 3.7).

Results

Learning on the Training Task

We calculated thresholds for the alpha level of the target gabor for each day of the training task. We modeled learning using a power function (Newell & Rosenbloom, 1981) to provide an estimate of the learning rate for each subject. A 1-sample t-test of the rate parameters for each subject indicated a significant learning rate ($t_5 = 5.62$, $p = 0.003$, see Figure 3.8). We can conclude that subjects in this task learn to locate the target object at lower alpha levels over time and that perceptual learning is occurring in this task.

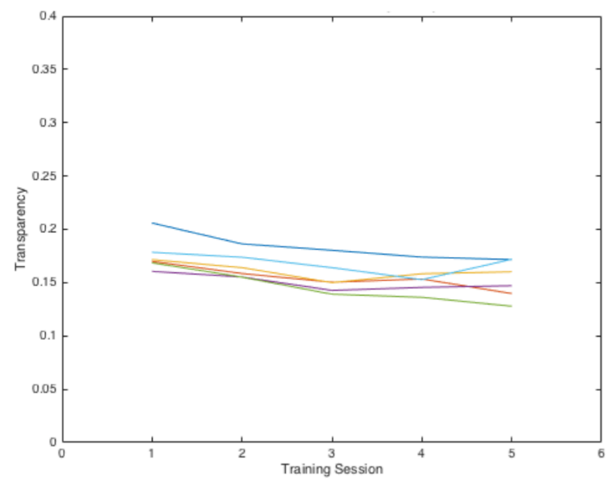


Figure 3.8. Mean transparency reached per session on each training day.

Pretest/Posttest Transfer Measures

As expected, at pre-test we found no effect of orientation on sensitivity to the target gabor (Figure 3.9A), but there was a significant effect of spatial frequency ($F_{2,30} = 18.00$, $p < 0.001$), and no interaction between the two. However, following visual search training there were significant main effects of both orientation ($F_{1,30} = 7.01$, $p = 0.013$) and spatial frequency ($F_{2,30} = 3.38$, $p = 0.047$), with a trend towards an interaction between the two ($F_{2,30} = 2.64$, $p = 0.088$, Figure 3.9B). Looking at both sessions together reveals the primary driver of the transfer effects; in addition to substantial main effects of spatial frequency ($F_{2,60} = 20.36$, $p < 0.001$) and testing day ($F_{1,60} = 14.37$, $p < 0.001$) we find a strong interaction between spatial frequency and day ($F_{2,60}$ = 5.05, p = 0.006) and a trend towards an interaction between orientation and frequency ($F_{1,60}$ = 3.29, p = 0.075, Figure 3.9C). These findings demonstrate that the visual search training generalizes to increased overall sensitivity to the visual components of the target gabor. However, the pattern of results suggests that the transfer effect of training was predominantly driven by improvements in overall familiarity with the stimulus and by sensitivity to lower spatial frequencies. Improvements specific to the trained stimulus appear to be focused on improved sensitivity to the target orientation, but these effects are weaker.

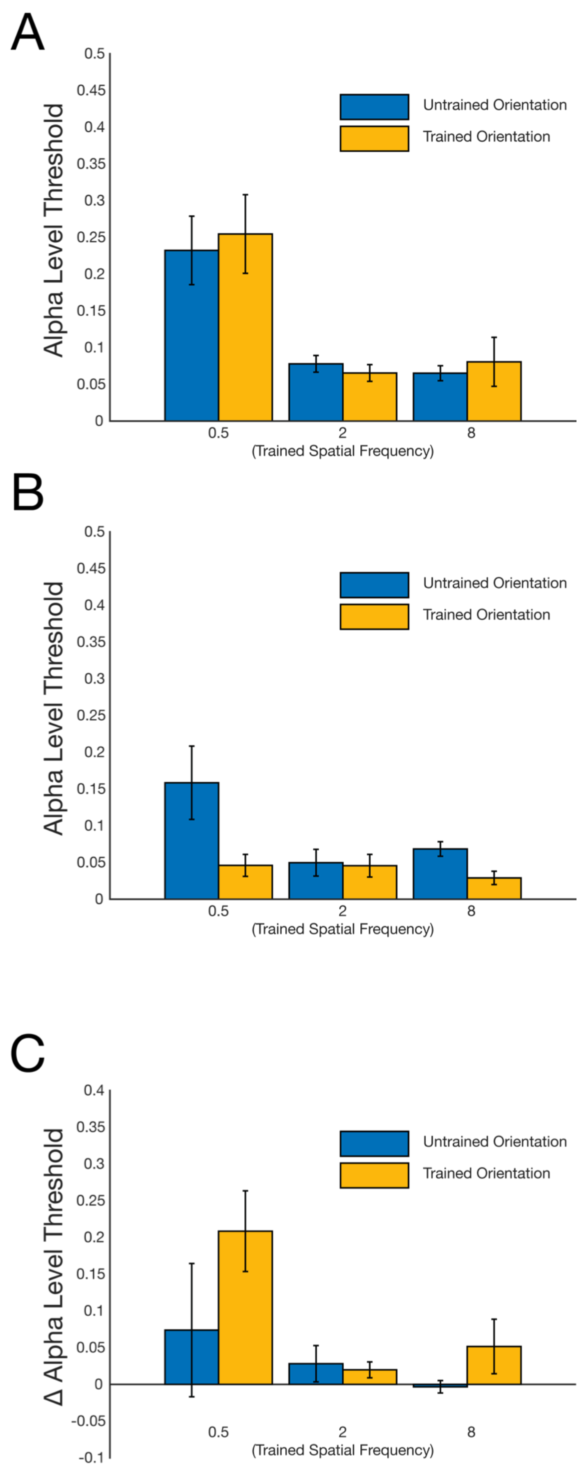


Figure 3.9. Mean pretest (A) and posttest (B) thresholds for the transfer task. Colored bars indicate the trained orientation, the groupings of bars indicate the spatial frequency of the gabor in cycles/degree. Difference scores for this measure are shown in part C.

We calculated two measures based on similarity score to illustrate the types of results this novel method can provide. First, we calculated the mean similarity scores for each non-target fixation, per trial. This reflects the ability of our approach to provide single-trial temporal resolution, but with reduced overall noise due to the combination of multiple non-target fixations per trial. See Figure 3.10 for a sample of one subject's average similarity scores over time in a single session. After calculating the mean similarity score per session, we fit a linear function to the scores and conducted a 1-way analysis of variance on the fitted slope parameters to determine if the average

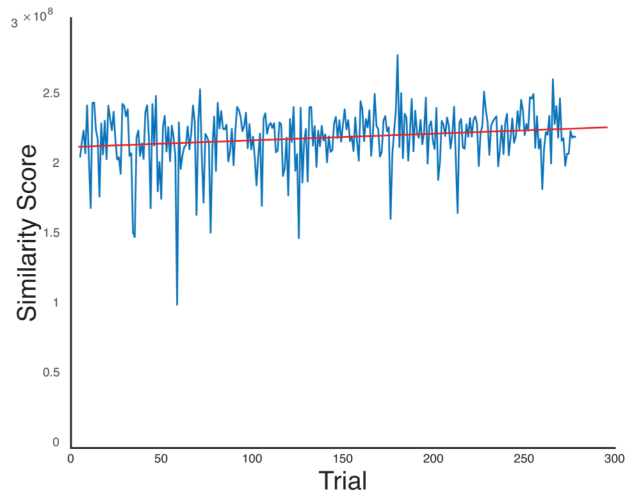


Figure 3.10. Mean similarity score for all non-target fixations, per trial, for one sample session for one subject. This figure illustrates the type of data the similarity score method can generate. Similarity scores are unitless, but higher values correspond to greater similarity. The red line indicates the linear best fit for the trend of scores in this session.

similarity scores were increasing over time across all our subjects. We found this to not be the case for our whole sample ($F_{4,25} = 0.85$, $p = 0.505$), although there is considerable between-subject variance in these results. Second, we calculated the ratio of dwell time by the similarity score for each non-target fixation, per trial. This reflects the degree to which subjects prefer to continue to look at portions of the image that contain energy which is similar to the target gabor. See Figure 3.11 for a sample of one subject's average similarity score ratio over time in a single session. As we did with the similarity scores, we fit a linear function to the ratios and conducted a 1-way analysis of variance on the fitted slopes to determine if the average ratios were increasing over time across all our subjects. We found this to not be the case for our whole sample ($F_{4,25} = 1.80$, $p = 0.161$), although as before there is considerable between-subject variance in these results. It is clear that the trial-by-trial data is still quite noisy, and we are unable to draw strong inferences about subject behavior from this sample. It is clear there is more work to be done to better understand these results. However, our primary purpose here is to demonstrate the potential utility of the data

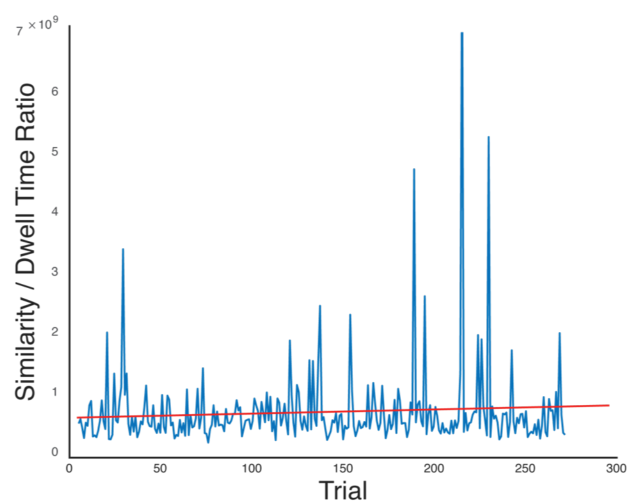


Figure 3.11. Mean ratio of dwell-time for each non-target fixation to the similarity score for that fixation, per trial, for one sample session for one subject. This figure is a unitless value over time, but higher values correspond to greater similarity. The red line indicates the linear best fit for the trend of score rations in this session.

analytic methods available to perceptual learning research thanks to the inclusion of eye tracking methods.

Finally, to illustrate the similarity between our new method and existing perceptual learning approaches, and to further demonstrate that perceptual learning is indeed occurring over the course of the visual search task, we looked at the total trial time for each subject per session. See Figure 3.12 for a sample of one subject's performance. These results include both correct and incorrect trials, but the observation of a negative slope for the average time corresponds to reduced time per trial over the course of a session when we consider the fact that the staircase procedure is maintaining a consistent level of correct and incorrect trials overall. As before, we fit a linear function to the trial lengths and conducted a 1-way analysis of variance on the slopes to determine if the average time was decreasing over time across all our subjects. We found a trend across our whole sample ($F_{4,25} = 2.46$, $p = 0.071$), suggesting that subjects are also completing the task more quickly overall in addition to detecting the target at a lower alpha level.

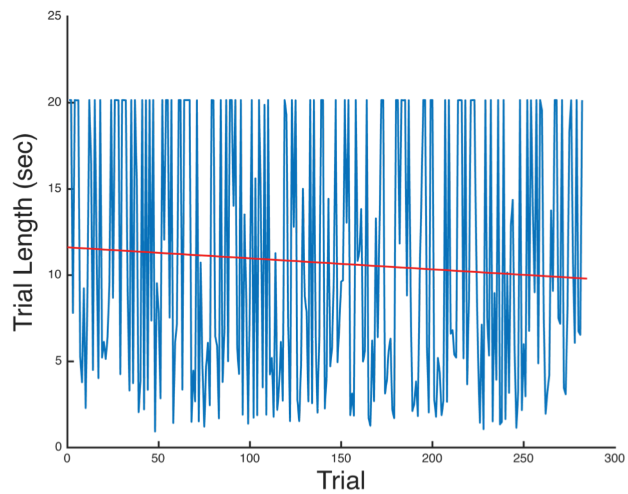


Figure 3.12. Length of each trial for one sample session for one subject. This figure corresponds to typical reaction time data, and is presented to illustrate the correspondence between the novel method we are proposing and existing perceptual learning methods. The red line indicates the linear best fit for the trend of times in this session.

Discussion

Experimental Discussion

The purpose of the empirical results in this article was to demonstrate the utility of our novel method for inducing and measuring perceptual learning. We have demonstrated that training with the visual search task does result in perceptual learning as well as transfer to other salient dimensions of the stimulus. We also presented both trial-by-trial and session-level results demonstrating measurements of behaviors relevant to perceptual learning. These results are only a small snapshot of the wide variety of potential measures that may be derived from such a rich dataset. Unfortunately, in our dataset some of the most interesting effects are overwhelmed by noise and so future work will be required to resolve these issues. Nevertheless, we hope that the results presented here and the unique measures we are able to utilize will be inspirational to other researchers who will further explore the frontiers of the dataspace that eye tracking-based methods can provide.

Method Discussion

The time course of learning is a critical area of inquiry for perceptual learning researchers. How quickly do subjects acquire the trained skill? During skill acquisition, which aspects of the stimulus are learned at what times? Is learning entirely implicit or do the strategies and conscious goals of the subject impact learning? Is it the case that important elements of perceptual learning occur within the first few trials with a new stimulus? All of these questions require a degree of temporal precision that is inaccessible using methods that depend upon either discrete testing sessions or large blocks of trials. Here we present a novel method that aims to provide the necessary temporal specificity to investigate these questions. Despite the fact that the present iteration of the method is too noisy for effective interpretation, we believe that this is a productive step in a positive new direction for improving temporal resolution in perceptual learning research. 3 decades have past since Fiorentini & Berardi (1980) plotted subject improvement over blocks of trials, and this basic approach has been used again (Tremblay, Kraus, McGee, 1998) and again (Atienza, Cantero & Dominguez-Martin, 2002); each time offering vital insights into the temporal dynamics of perceptual learning. It is time to take the next methodological step. Fiorentini & Berardi's (1980) study provided elegant proof of the specificity of perceptual learning to the trained orientation, and the rapid acquisition of new learning with exposure to novel orientations and novel spatial frequency. Tremblay, Kraus, & McGee (1998), made wonderful use of EEG techniques to measure rapid learning, however the indirect nature of neuroimaging techniques required them to merely speculate on what the underlying cognitive mechanisms may have been and they were obliged to fall back on theoretical ideas about fast perceptual learning (Polat & Sagi, 1995). And more recently Atienza, Cantero & Dominguez-Martin's (2002) extensive discussion of learning-related ERPs included an insightful discussion of the influence of attention on perceptual learning based on inferences from ERP components thought to correspond to attentional signals. However although they were able to discuss the differential onset of rapid and slow attentional processes and their impact on learning, they were unable to draw specific conclusions about the content of learning.

What are people actually learning when learning occurs? This question, while seemingly basic, is elusive in the field of perceptual learning. We believe that a method that would attempt to address this question must necessarily depend upon observed behaviors, rather than inferences from neuroimaging. In addition to the high cost of many neuroimaging techniques, all of these methods depend on strong inferential assumptions about the underlying relationship between the physical response measured and cognitive function, whereas this method requires only direct observation of subject behavior. We developed this approach as a step towards measuring the content of learning at a high temporal resolution and to provide researchers with an additional tool to examine important cognitive processes.

As part of validating this approach we needed to address several other research questions related to the utility of the task at large. First, can visual search training of this type induce perceptual learning? Clearly the answer is yes: subjects demonstrate reductions in detection threshold over time during the visual search task as well as measurable changes in sensitivity in the transfer task (See Figure 3.8). Second, can we use the data derived from our analytical techniques? Is the information interpretable? Certainly within-trial data based on single fixations contain substantial noise, and consequently we are unable to draw strong conclusions from them at this time. Nevertheless, we are optimistic that these issues can be resolved with time.

We present these preliminary results with this method to the perceptual learning community to inspire others to explore new research questions in new ways, and so we conclude by offering a few suggestions for possible research questions for which this approach may be informative. One possible change to the method is the inclusion of additional non-target lure objects with different visual properties from the target object. These lures can be used to draw inferences about specificity in the training task itself, as opposed to relying on measures of learning and transfer in the pre-test and post-test sessions (which are themselves a transfer measure compared to the visual search task). Other non-target objects may be employed as well, at one time we considered a task-irrelevant perceptual learning paradigm (Seitz & Watanabe, 2009) where subjects were instructed to search for a ring target, and the stimulus gabor was embedded below-threshold in the center of the ring. In private conversation with other researchers, we have even discussed applications using other noise filters to deliberately bias the types of information available to the subject. Novel research methods are often tumultuous in their beginnings, but although we were unable to show strong statistical results, future research using on this method and other methods which it may inspire will doubtless help expand our understanding of perceptual learning.

References

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

Andersson, R., Larsson, L., Holmqvist, K., Stridh, M., & Nyström, M. (2017). One algorithm to rule them all? An evaluation and discussion of ten eye movement event-detection algorithms. *Behavior Research Methods*, *49*(2), 616-637.

Atienza, M., Cantero, J. L., & Dominguez-Marin, E. (2002). The time course of neural changes underlying auditory perceptual learning. *Learning & Memory*, *9*(3), 138-150.

Ball, K., & Sekuler, R. (1987). Direction-specific improvement in motion discrimination. *Vision Research*.

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, 10, 433–436.

Censor, N., Sagi, D., & Cohen, L. G. (2012). Common mechanisms of human perceptual and motor learning. *Nature Reviews Neuroscience*, *13*(9), 658.

Donovan, I., Szpiro, S., & Carrasco, M. (2015). Exogenous attention facilitates location transfer of perceptual learning. *Journal of Vision*, *15*(10), 11-11.

Fahle, M., Edelman, S., & Poggio, T. (1995). Fast perceptual learning in hyperacuity. *Vision Research*, *35*(21), 3003-3013.

Fiorentini, A., & Berardi, N. (1980). Perceptual learning specific for orientation and spatial frequency. *Nature*, *287*(5777), 43.

Flege, J. E. (1987). A critical period for learning to pronounce foreign languages?. *Applied Linguistics*, *8*(2), 162-177.

Galliussi, J., Grzeczkowski, L., Gerbino, W., Herzog, M. H., & Bernardis, P. (2018). Is lack of attention necessary for task-irrelevant perceptual learning?. *Vision Research*, *152*, 118-125.

Gibson, E. J. (1963). Perceptual learning. *Annual Review of Psychology*, *14*(1), 29-56.

Green, D. M., & Swets, J. A. (1966). *Signal detection theory and psychophysics* (Vol. 1). New York: Wiley.

Hefferon, J. (2017). *Linear algebra*. Retrieved from http://joshua.smcvt.edu/linearalgebra/book.pdf.

Hessels, R. S., Niehorster, D. C., Kemner, C., & Hooge, I. T. (2017). Noise-robust fixation detection in eye movement data: Identification by two-means clustering (I2MC). *Behavior Research Methods*, *49*(5), 1802-1823.

Jacques, T., & Seitz, A. R. (2018, May). *Using eye tracking to develop classification images for perceptual learning*. Talk presented at the annual meeting of the Vision Sciences Society, St. Pete's Beach, FL.

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.

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.

Karni, A, & Sagi, D. (1991). Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(11), 4966–70.

Karni, A., & Sagi, D. (1993). The time course of learning a visual skill. *Nature*, *365*(6443), 250.

Leonards, U., Rettenbach, R., Nase, G., & Sireteanu, R. (2002). Perceptual learning of highly demanding visual search tasks. *Vision Research*, *42*(18), 2193-2204.

Lewis, C. M., Baldassarre, A., Committeri, G., Romani, G. L., & Corbetta, M. (2009). Learning sculpts the spontaneous activity of the resting human brain. *Proceedings of the National Academy of Sciences*, *106*(41), 17558-17563.

Lykins, A. D., Meana, M., & Kambe, G. (2006). Detection of differential viewing patterns to erotic and non-erotic stimuli using eye-tracking methodology. *Archives of Sexual Behavior*, *35*(5), 569-575.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Psychology press.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. *Cognitive skills and their acquisition*, *1*(1981), 1-55.

Nyström, M., & Holmqvist, K. (2010). An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data. *Behavior Research Methods*, *42*(1), 188-204.

Pärnamets, P., Johansson, P., Hall, L., Balkenius, C., Spivey, M. J., & Richardson, D. C. (2015). Biasing moral decisions by exploiting the dynamics of eye gaze. *Proceedings of the National Academy of Sciences*, *112*(13), 4170-4175.

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10, 437–442.

Polat, U., & Sagi, D. (1995). Plasticity of spatial interactions in early vision. In *Santa Fe Institute Studies in the Sciences of Complexity proceedings Volume 23.* Addison-Wesley Publishing Co.

Qu, Z., Song, Y., & Ding, Y. (2010). ERP evidence for distinct mechanisms of fast and slow visual perceptual learning. *Neuropsychologia*, *48*(6), 1869-1874.

Salvucci, D. D., & Goldberg, J. H. (2000, November). Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye Tracking Research & Applications* (pp. 71-78). ACM

Sasaki, Y., Nanez, J. E., & Watanabe, T. (2010). Advances in visual perceptual learning and plasticity. *Nature Reviews Neuroscience*, *11*(1), 53.

Seitz, A., & Watanabe, T. (2005). A unified model for perceptual learning. *Trends in Cognitive Sciences*, *9*(7), 329-334.

Seitz, A. R., & Watanabe, T. (2009). The phenomenon of task-irrelevant perceptual learning. *Vision Research*, *49*(21), 2604-2610.

Seitz, A. R. (2018). A new framework of design and continuous evaluation to improve brain training. *Journal of Cognitive Enhancement*, *2*(1), 78-87.

Shibata, K., Watanabe, T., Sasaki, Y., & Kawato, M. (2011). Perceptual learning incepted by decoded fMRI neurofeedback without stimulus presentation. *Science*, *334*(6061), 1413-1415.

Sireteanu, R., & Rettenbach, R. (1995). Perceptual learning in visual search: Fast, enduring, but non-specific. *Vision Research*, *35*(14), 2037-2043.

Tremblay, K., Kraus, N., & McGee, T. (1998). The time course of auditory perceptual learning: neurophysiological changes during speech-sound training. *Neuroreport*, *9*(16), 3557-3560.

Watson, C. S. (1980). Time course of auditory perceptual learning. *Annals of Otology, Rhinology & Laryngology*, 89(5), 96-102.

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.

Weiss, Y., Edelman, S., & Fahle, M. (1993). Models of perceptual learning in vernier hyperacuity. *Neural Computation*, *5*(5), 695-718.

Wichmann, F. A. & Hill, N. J. (2001). The psychometric function: I. Fitting, sampling, and goodness of fit. *Perception & Psychophysics*, *63*(8), 1293-1313.

Wiesel, T. N., & Hubel, D. H. (1965). Comparison of the effects of unilateral and bilateral eye closure on cortical unit responses in kittens. *Journal of Neurophysiology*, *28*(6), 1029-1040.

Wu, D. C., & Tsai, W. H. (2003). A steganographic method for images by pixel-value differencing. *Pattern Recognition Letters*, *24*(9-10), 1613-1626.

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.

Yotsumoto, Y., Watanabe, T., & Sasaki, Y. (2008). Different dynamics of performance and brain activation in the time course of perceptual learning. *Neuron*, *57*(6), 827-833.

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.

Zhang, P., Zhao, Y., Dosher, B. A., & Lu, Z. L. (2019). Assessing the detailed time course of perceptual sensitivity change in perceptual learning. *Journal of Vision*, *19*(5), 9-9.

General Discussion

The primarily methodological nature of this work represents an opportunity for perceptual learning researchers. As we discussed in the introduction to the dissertation, there are substantial practical applications for perceptual learning research ranging from therapeutic (Polat et al., 2009) to athletic (Deveau et al., 2014). However, we view the significance of the work to lie primarily in it's utility to researchers in the pursuit of their own applications and research questions.

In the process of conducting research it is often useful to take advantage of established methods and situate an experiment firmly within the existing literature. This practice helps improve the replicability of findings, helps resolve disputes between conflicting results, and provides inspiration for further research along the same lines. On the other hand, developing new techniques is the only way to establish new standards, and at times innovation is necessary to adequately characterize empirical results or to make new types of measurement.

In Chapter 1 we describe a traditional perceptual learning experiment based heavily on existing methods. The basic research design is not groundbreaking; we use a well-established perceptual learning task (Karni & Sagi, 1991) and closely duplicate even the smallest details of previous studies with this task (Yotsumoto et al., 2009). Its novelty lies chiefly in the large sample size and the ability to use that large sample to draw between-subjects inferences about the effects of pre-existing subject abilities, the effects of differential properties of the task procedure, and the interactions between the two on learning and transfer. The basic questions regarding the influence of visual attention skill and action video games on learning and transfer in the TDT represent a positive step towards better understanding how individual differences in ability influence perceptual learning, a topic which is greatly in need of continuing research. The inclusion of task-related effects is also an important contribution to a field that, in spite of the enormous impact of experimental procedures on learning outcomes, has yet to fully characterize these effects. And finally, this project also included a replication of its own: Green & Bavelier's (2003) reports of a strong relationship between action video game play and visual attention did not replicate in our population.

However, even a project as solidly grounded in established methods and past research as this was not immune to unexpected problems. Our finding that roving the order of trial difficulties led to reduced overall performance broke one of the major assumptions perceptual learning researchers make about low-performing subjects during data analysis. These results obliged us to to deviate from traditional data analytic techniques because low performing subjects could no longer be excluded from the sample offhand. In Chapter 2 we explore the profound impact even this small break from established methods had on our ability to evaluate our results, and in response we broke new ground in understanding what it means to be a valid subject in perceptual learning tasks. We innovated, developing a unique approach to the basic and age-old practice of dividing participants using a median split. And armed with our new approach we demonstrated that our original findings in the TDT were indeed valid, and we can add these new techniques to the toolbox of methods available to perceptual learning researchers.

Finally, in Chapter 3 we broke fully from established methods to create a completely new paradigm for understanding perceptual learning. The oldest methods in cognition are behaviorally based, and observing human behavior directly is still one of the most common ways to draw inferences about mental processes. Eye tracking is a powerful tool for observing behavior, not only giving us strong insight into overt shifts of attention during a task but also by providing high temporal resolution and by allowing secondary measurements of behavior such as pupilometry. By marrying this technology with a free-form visual search task and carefully selected stimuli we are able to open up new avenues of inquiry for perceptual learning researchers. Although this technique still requires additional validation and refinement, this method opens the doors for a wide variety of experimental questions which were previously inaccessible or difficult to measure.

Conclusion

Although the concept is likely much older, the phrase “if all you have is a hammer, everything looks like a nail“ is credited to Maslow’s (1966) *Psychology of Science*. The tools and methods available to researchers define and can limit our understanding of the world and impose a type of intellectual myopia in studying it. Certainly, established methods are the “gold standard” for a good reason, but it is important to develop new methods when necessary and appropriate. With this work, we contribute to our understanding of perceptual learning in general, and also to the methods cognitive psychologists employ to better understand learning.

The methodological innovations we have described can be used to further our understanding of many aspects of human learning behavior. Certainly niche cases where a single criterion based on a joint condition is necessary can benefit from the approach described in Chapter Two, but we are particularly enthusiastic about the opportunities presented by the eye tracking method in Chapter Three. Direct data regarding eye movements have already been used in a variety of scene perception applications including visual search (see Rayner, 2009, for a discussion of the topic), as well as more direct learning applications such as face learning (Henderson et al., 2005) and multimedia learning (Mayer, 2010). However, these approaches rely primarily on superficial aspects of gaze data such as image content, rather than on the underlying statistics of the image, a distinction that may be more productive from a perceptual perspective. A deeper understanding of what aspects of visual information are relevant to learners at different phases of the learning process could potentially revolutionize our understanding of training in professions where visual expertise is critical such as radiology, military drone pilots, and security screeners. The direct applications of this approach on visual task learning are clear, however the method in Chapter Three has been deliberately described in a generalized fashion in order to maximize the potential for applications to questions we cannot foresee at this time. A principled approach to interpreting what are essentially false alarms in visual search, and exploiting that information for scientific benefit has research potential far beyond perceptual learning alone. We eagerly anticipate the discoveries that will certainly come from creative application of the ideas that support this technique.

References

Deveau, J., Ozer, D. J., & Seitz, A. R. (2014). Improved vision and on-field performance in baseball through perceptual learning. *Current Biology*, *24*(4), R146-R147.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*(6939), 534–7.

Henderson, J. M., Williams, C. C., & Falk, R. J. (2005). Eye movements are functional during face learning. *Memory & cognition*, *33*(1), 98-106.

Karni, A, & Sagi, D. (1991). Where practice makes perfect in texture discrimination: evidence for primary visual cortex plasticity. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(11), 4966–70.

Maslow, A. H. (1966). The psychology of science: A reconnaissance. New York: Harper & Row.

Mayer, R. E. (2010). Unique contributions of eye-tracking research to the study of learning with graphics. *Learning and instruction*, *20*(2), 167-171.

Polat, U., Ma-Naim, T., & Spierer, A. (2009). Treatment of children with amblyopia by perceptual learning. *Vision research*, *49*(21), 2599-2603.

Rayner, K. (2009). Eye movements and attention in reading, scene perception, and visual search. *The quarterly journal of experimental psychology*, *62*(8), 1457-1506.

Yotsumoto, Y., Chang, L. H., Watanabe, T., & Sasaki, Y. (2009). Interference and feature specificity in visual perceptual learning. *Vision Research*, *49*(21), 2611–2623.