From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Psychology

2020

A systematic investigation of prediction error and learning in infancy

Zhang, Yue

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp014j03d2576

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

A SYSTEMATIC INVESTIGATION OF PREDICTION ERROR AND LEARNING IN INFANCY

Yue (Felicia) Zhang

A DISSERTATION

PRESENTED TO THE FACULTY

OF PRINCETON UNIVERSITY

IN CANDIDACY FOR THE DEGREE

OF DOCTOR OF PHILOSOPHY

RECOMMENDED FOR ACCEPTANCE BY

THE DEPARTMENT OF PSYCHOLOGY

Advisers: Lauren Emberson, Casey Lew-William

June 2020

© Copyright by Yue Zhang, 2020. All rights reserved

Abstract

Prediction, or the ability to use past experiences to generate expectations about future sensory input, is increasingly believed to be crucial for adult cognition and learning. Despite decades of targeted research on prediction in adults, it remains unclear about how prediction operates during infancy. In Chapter I, I discuss how pupillometry, or the measurement of pupil size, can be used to investigate prediction and its component parts, namely anticipation, expectation, prediction error and updating, with an emphasis on prediction error. In Chapter II, I take a behavioral (measuring infants' pupil size) and computational approach (using an error-driven learning model, the Rescorla-Wagner Model) to understand how infants use prediction error and whether this ability changes across the lifespan. Chapter III describes a series of studies that uses pupil size to measure prediction in a standard visual statistical learning task. Chapter IV sought to combine eye-tracking and pupillometry in order to understand how prediction error impacts infants' ability to predict, or elicit anticipatory eye movements, in the context of the Visual Expectation Paradigm. Lastly, Chapter V is a follow-up investigation that also uses eye-tracking and pupillometry, to determine whether prediction error is the main driver behind infants' ability to update. Taken together, these studies expand on the theories that have proposed that prediction is a core aspect of adult brain function to human development and specifically investigates how prediction and prediction error supports learning in infancy.

Acknowledgements

I am truly grateful for all the wonderful people who accompanied me along this journey. This dissertation would not have been possible without the support of each and every one of you. First, thank you to my advisor, Lauren Emberson, for providing me with exceptional mentorship and guidance. You have taught me how to approach science with more rigor, optimism and confidence. Thank you for being patient and always going above and beyond with providing feedback on my drafts – I am a better writer thanks to you. Thank you for creating such an intellectually stimulating and supportive lab environment, I owe much of my development over the past 5 years to you.

Thank you to my secondary advisor, Casey Lew-Williams. I am so lucky to have worked with you not only in the lab, but also in the classroom. Thank you for providing me with invaluable teaching moments and giving me the freedom to explore and develop my own teaching methods. Watching you teach and lead has been extremely inspirational.

Thank you to Jordan Taylor for always asking thoughtful questions and helping me think about my research from a different perspective.

Thank you to my other committee members: Yael Niv for your instrumental feedback, Tania Lombrozo for your time and Jon Cohen for your input.

Thank you to my former mentors who played a huge role in getting me to where I am today: Morgan Barense, Jonathan Erez, Joyce Mak, Jennifer Hardy, and Nasreen Khatri.

Thank you to all the Baby Lab members that shaped my graduate school experience. Tracy Reuter, I am so happy to have started and end this journey with you. It's been so fun working with you on different projects and I am so grateful for our friendship. Jessie Schwab, thank you for introducing me to so many new opportunities and for mentoring me during periods

of confusion, I can't thank you enough. Thank you Sagi Jaffe-Dax and Gabriel Xiao for your support and for walking me through things I don't understand. Claire Robertson, thank you for being you, I miss your presence in the lab. Thank you Yaelan Jung and Annie Schwartzstein for making my last year of graduate school more special. Thank you Sori Baek and Anna Herbolzheimer for being awesome office mates and the best distractions. Thank you to Alex Boldin and Carolyn Mazzei for the fun early years of Baby Lab and thank you to Elise Piazza, Sammy Floyd, Karina Tachihara, Mira Nencheva, Juliana Trach, Chris Potter, Crystal Lee, Jessica Kosie for making the Baby Lab a truly special place. Importantly, thank you to all of the research assistants that supported this research, especially Wen Du, Alice Wang and Sam Einspahr. Sam, thank you for all your hard work, I can't wait to see what you do in the future!

I am lucky to have met so many amazing people throughout graduate school. Thank you to Na Yeon Kim, Joseph Avery, Aaron Kurosu, Zidong Zhao, Judith Mildner and Victoria Ritvo for fun times working together on non-research related projects and chatting about life. I appreciate each and every one of you.

Thank you to my dear friends for making my graduate school experience more fun and meaningful. Simon Quach, we met before graduate school even started and your constant uplifting energy helped me through so much. Lindsey Eikenburg, thank you for never failing to bring up my pronunciation of "Dulce de Leche" and for sticking it out with me in the GC for so long. Karolina Koziol, thank you for introducing me to all the delicious vegetarian food, and I can't wait for more New York adventures together. Gandalf Nicolas, I am so glad to have you as a friend. You are hilarious, you are brilliant, and I am so excited to follow your academic journey. Thank you Monica Stanton for being the most supportive friend despite living on the other side of the country. You have taught me how to enjoy the little things in life and introduced

me to so many new things that I can't even begin to list them all. I am so grateful for your friendship.

To my loved ones, these 5 years would not have been possible without you.

Thank you Coffee Meets Bagel for connecting me with Wes. And thank you Wes. I love you – you are a strange one. You make me laugh and inspire me every single day with your dedication to better humanity. Also, thank you for keeping me sane during these crazy times. There's no one else I would rather social distance with.

Thank you mom and dad, Hongjie and Mei, for believing in me and teaching me the importance of hard work. You two have sacrificed so much for us to be here and I am eternally grateful to have you as parents. Thank you to my grandparents for their support from afar. Thank you Alice, my little sister, my best friend, for constantly making me laugh and for sometimes playing the big sister role; I love you all so much.

Table of Contents

Abstract……...……iii

Acknowledgments…….iv

List of Tables and Figures……..…ix

Introduction…….……1

Chapter I: Using pupillometry to investigate predictive processes in infancy… … … … … … …8

Chapter II: Prediction in infants and adults: A pupillometry study (Experiment 1)...……..…34

Methods……..38

Results……42

Discussion……...……47

Chapter III: Opposing timing constraints severely limit the use of pupillometry to investigate visual statistical learning……..……..51

General methods……..……..58

Experiment 2: SL with triplets and two-alternative forced-choice test……...…….61

Methods…….……61

Results……64

Discussion……..……68

Experiment 3: SL with paired images and recognition test……...……70

Methods……..……70

Results……71

Discussion……..……76

Experiment 4: VSL using triplet and random blocks…….……..……..77

Methods…….…….……78

Results……78

Discussion……..……..……81

General discussion……...……..81

Chapter IV: Exploring prediction and prediction error using eye movements, pupillometry and learning models (Experiment 5)……..……90

Methods……98

Results……..……104

Discussion……117

Chapter V: Is prediction error the mechanism behind infants' ability to update anticipation?...122

Experiment 6: Infants……...…127

Methods……128

Results……..……132

Discussion……139

Experiment 7: Adults……....…139

Methods……140

Results……140

Discussion……145

General discussion……146

Conclusion……..151

References…….....157

Appendices……..……171

List of Tables and Figures

Figure 1. Example of how mean looking time and pupil diameter can be used in the same study..13

Figure 2. Pupil change for visual omission and visual present trials...……25

Figure 3. Comparing simulated parameters generated from the Rescorla-Wagner model for adults and infants on a trial-by-trial basis..…….…..…….……26

Figure 4. Trial procedure……39

Figure 5. PDR plotted against time for the two trial types…..…….. 43

Figure 6. Scatterplot of average PDR (average calculated from immediately after the visual stimulus appears/supposed to appear to 4000ms) to present trial versus omission trial, for both age groups…….……44

Figure 7. Correlation of PDR and prediction error magnitude (δ^2) from the model for every trial of every subject to determine model fit..…….…..…….……45

Figure 8. Comparing parameters from the learning model for infants and adults……46

Figure 9. Comparing simulated parameters generated from the infant and adult learning model on a trial-by-trial basis…….…..…….……47

Figure 10. The formation of a trial in a standard VSL study compared to a trial in the current study…….…..…….……..55

Figure 11. The 12 fractal images used in Experiment 2, Experiment 3 and Experiment 4…..….59

Figure 12. Schematic of VSL task in Experiment 2..…….…….……...….62

Figure 13. Distribution of test phase accuracy in Experiment 2…….…….……...….65

Figure 14. Pupil timecourse during viewing of a triplet in Experiment 2…….……...….66

Figure 15. Distribution of test phase accuracy in Experiment 3……...…..……...….72

Figure 16. Pupil timecourse in Experiment 3 using the anticipatory period of the first image in a pair..…..……...…….…...74

Figure 17. Pupil timecourse in Experiment 3 using the anticipatory period of the most recent unpaired image as baseline…..……...75

Figure 18. Pupil time-course during viewing of a triplet in Experiment 4……...79

Table 1. Overview of the two different baseline selection results for all experiments...……...82

Figure 19. Changes in pupil diameter based on X coordinate eye position for left eye (left panel) and right eye (right panel)……..……...86

Figure 20. Schematic of the modified Visual Expectation Paradigm during the test phase…...97

Figure 21. Schematic of trial…..……..101

Figure 22. Proportion of AEMs (the number of correct AEMs / total number of valid trials) as a function of sequence (Semi-Predictable and Predictable)……..…….....106

Figure 23. Median REM RT as a function of sequence (Semi-Predictable and Predictable) and its respective baseline..…..…….……...108

Figure 24. Median REM RT as a function of target (Predictable Target from the Predictable Sequence, and Common Target and Prediction Error Target from the Semi-Predictable Sequence)..…..……..…..109

Figure 25. Average pupil change during when infants are looking at the target on the screen (1200-2200ms), grouped by sequence (Predictable and Semi-Predictable)……110

Figure 26. Average pupil change during when infants are looking at the target on the screen (1200-2200ms), grouped by target (Predictable Target from the Predictable Sequence, and Common Target and Prediction Error Target from the Semi-Predictable Sequence)……..……111

Figure 27. Initial Parameters for the (A) RW Model and (B) PH Model……...……112

Figure 28. Trial-by-trial changes in model parameters, Prediction and Prediction Error, generated separately for each sequence for the (A) RW Model (B) PH Model……..……114

Figure 29. Trial by trial changes in model parameters, Prediction and Prediction Error, generated separately for each sequence using ideal initial parameter values for the (A) RW Model (B) PH Model..…..……..…....115

Figure 30. Schematic of one block in the experiment……..…….130

Figure 31. Proportion of trials that had correct or incorrect AEMs, separated by pre-switch and post-switch for infants..…..……..……..…...133

Figure 32. Proportion of infants that generated a correct AEM versus incorrect AEM on a trial by trial basis during post-switch when the target appeared in the novel location……..…..134

Figure 33. Average pupil change during target period for pre-switch versus post-switch trials for infants..…..……..……..136

Figure 34. Time-course of pupil change for pre-switch trials and post-switch trials for infants..…..……..…..…...137

Figure 35. Average pupil change during target period for pre-switch trial 8 versus post-switch trial 1 for infants……..……..……..…138

Figure 36. Average pupil change during target period for pre-switch trial 8 versus post-switch trial 1 for each block for infants..…….……..…...139

Figure 37. Proportion of trials that had correct or incorrect AEMs, separated by pre-switch and post-switch for adults..……..……..…...141

Figure 38. Proportion of adults that generated a correct AEM versus incorrect AEM on a trial by trial basis during post-switch when the target appeared in the novel location……142

Figure 39. Average pupil change during target period for pre-switch trial 8 versus post-switch trial 1 for adults……..……..……..…...143

Figure 40. Timecourse of pupil change for pre-switch trial 8 and post-switch trial 1 for adults….....……..……..…...144

Figure 41. Average pupil change during target period for pre-switch trial 8 versus post-switch trial 1 for each block for adults…….……...……..…...145

Introduction

Learning plays an important role in human cognition: Why we behave and think as we do can be traced back to information learned from our environment. Especially relevant for developmental psychology is how infants learn so much and so efficiently during their first year of life. For example, before the age of 12 months old, infants are linking common words to their referents (Bergelson & Swingley, 2012), demonstrating preference for musical patterns that are common in their environment (Hannon & Trainor, 2007; Trainor & Trehub, 1992) and predicting the goal of different actions (*eating*: Gredebäck, Lindskog, Juvrud, Green & Marciszko, 2018; Kochukhova & Gredebäck, 2010; *reaching*: Kanakogi & Itakura, 2011). In addition, newborns and young infants are sensitive to the statistics in their environment regardless of the modality (*auditory*: Saffran, Aslin, & Newport, 1996; Saffran, Johnson, Aslin, & Newport, 1999, *visual*: Bulf, Johnson, & Valenza, 2011; Fiser & Aslin, 2001; Kirkham, Slemmer, Richardson, & Johnson, 2007; Kirkham, Slemmer, & Johnson, 2002).

Although research has uncovered a lot about what information infants can learn from, there is still much to discover about *how* infants learn. This dissertation focuses on one possible learning mechanism: prediction, or the ability to use past experiences to generate expectations about future sensory input. Prediction can be correct or incorrect, and each has the potential to support incremental cognitive changes. Research with adults shows that when your prediction is correct, it allows you to use what you have learned to support more efficient processing of your sensory input (e.g. quickly interpret noisy or ambiguous stimuli, Kersten & Yuille, 2003; Kutas, DeLong & Smith, 2011; Wei & Stocker, 2015). When your prediction is incorrect, this can generate prediction error signals in the brain (Friston, 2005; Hollerman & Schultz, 1998; McClure, Berns & Montague, 2003; Rao & Ballard, 1999), which initiates the updating of our

prediction based on new information (i.e. learn). Put simply, prediction error can be the spark that initiates learning.

Despite the large role prediction error plays in adult learning (Koenig, Uengoer & Lachnit, 2018; den Ouden, Friston, Daw, McIntosh & Stephan, 2009; Hollerman & Schultz, 1998; Pearce & Hall, 1980; Rescorla & Wagner, 1972; Wills, Lavric, Croft & Hodgson, 2007), there is limited work directly relating prediction error to learning in infancy. One likely reason behind the lack of studies examining prediction error and learning in infancy is methodological constraints. The majority of studies that seek to understand infant learning often use offline measures or outcome measures, which are collected after learning has occurred and easier to collect with infants (Aslin, 2007; Karuza, Emberson & Aslin, 2014). For example, in standard statistical learning studies, infants view a stream of stimuli composed based on a statistical rule, and evidence for infants learning the statistical rule is collected during a test phase, which is implemented after infants viewing the stimuli (Bulf et al., 2011; Fiser & Aslin, 2001; Kirkham et al., 2002; Kirkham et al., 2007; Saffran et al., 1996; Saffran et al., 1999). In addition, measures such as saccades and fixations, which can be collected during the learning process, are often compiled into high-level summary statistics (Aslin, 2012), thereby losing the ability to probe how learning unfolds on a moment by moment basis. Overall, due to the prevalence of offline/outcome measures and summary statistics in developmental psychology, much less is known about how prediction error operates in infancy.

Fortunately, there is an infant friendly, non-invasive measure, that is already collected by eye-trackers, that can mitigate these methodological constraints: pupil size. There are several advantages to using pupillometry, or the measurement of pupil size, with infants. First, pupillometry allows for recording of the same physiological response across different ages in the

same tasks (Ross-Sheehy & Eschman, 2019; Wetzel, Buttelmann, Schieler & Andreas 2016; Zhang, Jaffe-Dax, Wilson, & Emberson, 2018). If we use similar tasks across populations, we can uncover the development of these cognitive processes. Second, pupillometry is meaningful at each time point, which allows you to exploit the high temporal resolution of eye tracking to examine how cognitive processes unfold over time (similar to a method like electroencephalogram (EEG)). In addition, pupillometry can also be used in conjunction with other measures such as looking time and eye movements (Cheng, Kaldy, & Blaser, 2019; Kleberg, del Bianco, & Falck-Ytter, 2019; Ross-Sheehy & Eschman, 2019; Tummeltshammer, Feldman & Amso, 2019). Finally, pupillometry can be used to bridge developmental psychology with computational modelling approaches, providing new and important insights about infant cognition. All of these cases demonstrate how pupillometry can be used to not only provide new insights about prediction error, but learning in general, in infancy.

As mentioned earlier, one benefit of using pupillometry is the ability to incorporate computational modelling with pupil data. This unique combination was recently explored in an adult pupillometry study (Koenig et al., 2018). The researchers used two learning models that clearly map out the relationship between prediction and prediction error (Koenig et al., 2018). First, the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972) is an associative learning theory that describes how the association between a cue and stimulus is formed. This theory states that the changes in association is directly determined by the prediction error. Second, the Pearce and Hall model (1980), another associative learning theory, states that prediction error affects the attention given to the stimuli, thereby indirectly changing the association between cue and stimulus. Using a series of three experiments, Koenig et al. (2018) determined that pupil change across the three tasks followed the same pattern as predicted by the Pearce and Hall

model. Overall, the significance of the two learning theories is they mathematically describe how prediction error is generated and used to update one's prediction or association, which allows researchers to quantify and assess prediction error more precisely. Moreover, building on the success of the studies conducted by Koenig et al. (2018), the same approach can be applied to developmental research.

Given the current gap in the literature involving a limited understanding of how prediction error operates during infancy, and the potential of using pupillometry and computational models in filling this gap, this dissertation will present some of the first developmental studies that sought to combine pupillometry with computational learning models. Ultimately, incorporating pupillometry as a method in developmental psychology extends yet complements existing methods and has the promise to expand our understanding of human cognition. This work expands on the theories that have proposed that prediction is a core aspect of adult brain function to human development and investigates if prediction error is a powerful learning mechanism available to support learning in infancy.

Chapter 1 provides an overview of research that uses pupillometry to study predictive processing. We include relevant adult and infant pupillometry studies that will facilitate infancy researchers to adopt this technique. We argue that there are two specific features of pupillometry that make it particularly useful for augmenting the investigation of prediction in infancy. One, pupillometry has excellent temporal resolution that will facilitate the differentiation of subcomponents of prediction (i.e., anticipation, expectation, prediction error, updating). Two, pupillometry is highly continuous across the lifespan, allowing researchers to directly compare responses between infants and adults using an identical paradigm: A rare methodological

possibility for infancy researchers and one that will greatly facilitate future research. A version of this chapter is currently under review at *Infancy*.

Chapter 2 describes a study that sought to determine whether prediction is an ability that is continuous across the lifespan or whether an infant's ability to predict is different from an adult's, qualitatively or quantitatively. We used pupillometry to provide the first direct comparison of prediction abilities across these disparate age groups. We measured pupil size in 6-month-old infants and adults as they completed an identical implicit learning task designed to help learn associations between sounds and pictures. We found significantly larger pupil change for visual omission trials (i.e. trials that violated participants' predictions) compared to visual present trials (i.e. trials that confirmed participants' predictions) in both age groups. Furthermore, we fitted the Rescorla-Wagner model to the pupil data and found similar time-course across age groups suggesting a continuity of predictive capacity and learning across these ages. This chapter was published as an article in *Developmental Science* (Zhang et al., 2018).

Chapter 3 is a compilation of three studies that investigates whether pupillometry is a suitable online measure (a measure that is collected during the learning process) for visual statistical learning (VSL) tasks, where performance is often investigated based on offline measures (measures collected after the learning is completed). While the properties of pupil size, such as its high temporal resolution and its relationship to prediction, have the potential to expand studies in VSL, through a series of three experiments, we find that the two are not compatible with each other. Our results revealed that pupil size, when used to index prediction, is not related to offline measures of learning. We also find pupil differences that appear to be a result of prediction, are actually a result of where we chose to baseline instead. Ultimately, we conclude that the fast-paced nature of VSL paradigms make it incompatible with the slow nature

of pupil change and pupillometry should not be used to investigate learning mechanisms in VSL. This chapter was published as an article in *Frontiers in Psychology* (Zhang & Emberson, 2019).

Chapter 4 sought to examine prediction, prediction error and learning, through collecting eye movements and pupil size during a modified Visual Expectation Paradigm (VExP, Haith, Gazan & Goodman, 1988). A standard VExP is composed of visual targets appearing in a left-right alternating sequence, where pictures on one side (e.g., left) are predictable (predictable sequence) and pictures on the other side (e.g., right) are random (unpredictable sequence). These studies found infants initiated anticipatory eye movements (AEMs) for the predictable sequence but not the unpredictable sequence. Furthermore, when infants failed to make an AEM, they produced a faster reactive eye movement (REM) to the target after it appeared. Improvements in both correct AEM and REM speed are considered evidence for infants learning the pattern in the predictable sequence and forming predictions. In the current study, we tested 8-month-old infants on a modified VExP, where we replaced the unpredictable sequence with a semi-predictable sequence, designed to elicit prediction error. Results revealed no significant findings when we compared proportion of AEMs, speed of REMs or pupil change between the Predictable and Semi-Predictable Sequence. We believe these null findings are a manifestation of methodological challenges that arise when combining eye movements and pupillometry, trial duration and modifying the standard VExP. We conclude this chapter by laying out the challenges that researchers should be aware of when combining eye movements and pupillometry and potential solutions.

Chapter 5 is a follow-up study that measures eye movements and pupillometry in the same task, to determine whether prediction error is the main driver behind infants' ability to update. The hypothesized link between prediction error and infants' ability to update is

theoretically and computationally sound, yet there are few behavioral studies that investigate this. In this chapter, we conducted a follow-up study of Reuter, Emberson, Romberg & Lew-Williams et al. (2018), where they found infants who updated predictions more readily (generate more correct AEMs) learned language better (have a larger vocabulary), but the mechanism that supports infants' ability to update prediction remains unknown. Thus, we tested infants and adults in an extended version of the task used in Reuter et al. (2018), and included pupil size as a measure of prediction error. We examined how pupil size during trials when the target switches to a novel location, which should generate prediction error, relate to the proportion of correct AEMs. Despite finding eye movements that suggest infants and adults are making predictions and performing the task properly, pupil size was not reflective of prediction error in this task. We discuss these results in the context of the paradigm, and similar to Chapter 4, discuss the extra steps that must be taken when measuring pupil size and future directions in this area of research.

Despite many important discoveries regarding what infants learn from different types of input, we have a great deal to uncover about the underlying mechanism of learning in infancy (i.e. how infants learn). My dissertation investigates the online mechanisms supporting learning in infancy by considering prediction and prediction error as candidate learning mechanisms. Moreover, I employ novel combinations of pupillometry, eye movements and/or computational learning models in my investigations.

Chapter I: Using pupillometry to investigate predictive processes in infancy[1]

Prediction, or the ability to use past experiences to generate expectations about future sensory input, is believed to support learning by connecting what has been learned in the past with what one is currently experiencing. While the importance of prediction is well-supported based on studies with adults, little direct research has investigated prediction in infancy. Research in this area is hampered because there are methodological limitations to studying prediction with infants. In this article, we argue that pupillometry, or the measurement of pupil size, has the potential to be an effective method to examine predictive processes in infants and to compare prediction across developmental stages. In this article, we summarize research using pupillometry to measure prediction in adults and recent findings in the use of pupillometry in infants. Ultimately, this review will give researchers an overview of how pupillometry can be used to investigate the subcomponents of prediction (anticipation, expectation, prediction error and updating), and the broad applicability of pupillometry as a tool for developmental research. We start with an overview of the physiology of the pupil and the established links between changes in pupil size and changes in cognition.

Pupillometry: Brief Overview

The pupil is the circular opening in the iris that allows light into the eye. Pupil size increases in darker environments to allow more light in and decreases in brighter environments. This automatic adjustment of pupil size based on light level is known as the pupillary light reflex and is important for achieving optimal vision across many environments.

---

[1] A version of this chapter is currently under review. Zhang, F. & Emberson, L. L. (under review). Using pupillometry to investigate predictive processes in infancy.

In addition to being modulated by light levels, the pupil is also (subtly) modulated by cognition; therefore, pupillometry, or the measurement of pupil size, has been long embraced by psychology research. Before the invention of eye-trackers, pupil size was measured manually based on images taken of the pupil. These early practices were extremely time-consuming but were the first studies to reveal a link between cognition and pupil size (Hess & Polt, 1964; Kahneman & Beatty, 1966). With the invention of eye-trackers, pupil size is now automatically recorded at high temporal resolution. The relative ease of this method has generated a renewed interest in using pupillometry as a measure in cognitive psychology.

Pupil size is modulated by cognition through the actions of a brain region called the locus coeruleus (LC). This subcortical brain structure is the main producer of norepinephrine (NE), an important neurotransmitter in the central nervous system involved in attentional control (Aston-Jones & Cohen, 2005), memory retrieval (Sterpenich et al., 2006), and general arousal (Samuels & Szabadi, 2008). Especially relevant for pupil size is the finding that neurons in the LC-NE system (and correspondingly NE) have two modes: **tonic** and **phasic** (Aston-Jones & Cohen, 2005; Cohen, McClure, & Yu, 2007), and these two modes have been found to correspond to differences in information-seeking behaviors[2]. As a result, tonic and phasic pupil change also reveal distinct information about how humans process information.

[2] Differences in information-seeking behaviors with respect to tonic and phasic mode are known as explore and exploit, respectively. The tonic mode leads to an increase in distractibility and is referred to as the exploration mode. The phasic mode leads to focused attention and optimized performance on a task and is referred to as the exploitation mode. For a detailed review about these two modes of activity, see Aston-Jones and Cohen (2005).

With respect to behavioral research, tonic mode is analogous to pupil measurements that are collected continuously over a longer period of time (most of the time as a baseline measurement). For adults, tonic pupil size or baseline pupil size has been measured to study intelligence (Tsukahara, Harrison, & Engle, 2016), risk taking behavior (Yechiam & Telpaz, 2011), motivation (Wykowska, Anderl, Schubö, & Hommel, 2013) and control state (Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010). However, tonic pupil size is not infant-friendly due to the fact that tonic/baseline pupil size requires a long period of time with little or no stimulation. Therefore, the current review focuses on pupillometry studies that use phasic pupil change and how phasic pupil responses can be used to investigate prediction in infancy.

The phasic mode is analogous to time-locked or stimulus-locked pupil measurements (e.g., the change in pupil size from baseline in response to the presentation of a stimulus). For adults, phasic pupil change has been used to investigate attention (Mathôt, van der Linden, Grainger, & Vitu, 2013), cognitive load (review: van der Wel & van Steenbergen, 2018), memory (review: Goldinger & Papesh, 2012), uncertainty (Geng, Blumenfeld, Tyson, & Minzenberg, 2015; Lavín, San Martín, & Rosales Jubal, 2014), and prediction, the main focus of this review (Koenig, Uengoer, & Lachnit, 2018; Nassar et al., 2012; O'Reilly et al., 2013; Zhang, Jaffe-Dax, Wilson, & Emberson, 2018). To date, infant pupillometry studies have used phasic pupil change to investigate numerous domains in infancy such as emotion (Geangu, Hauf, Bhardwaj, & Bentz, 2011; Hepach & Westermann, 2013) and language development (Byers-Heinlein, Morin-Lessard, & Lew-williams, 2017; Kuipers & Thierry, 2013; Tamási, McKean, Gafos, Fritzsche, & Höhle, 2016).

Overall, research with pupillometry is less common for infants than adults due to methodological challenges. For example, pupillometry requires participants to stay relatively still in order to get a precise measurement of pupil size. Furthermore, infant studies rely on attention-grabbing videos and realistic-looking images to maintain infant attention. If these images are not rendered isoluminant (i.e., have the same light levels regardless of picture), these images, at best, contribute noise and, at worst, confound the measurement of the pupil. However, if these images are rendered isoluminant, it can result in highly-distorted images and videos that are often less engaging or recognizable to infants. These and many other methodological challenges can be overcome; Hepach and Westermann (2016) provide an excellent overview of practical considerations of infant pupillometry studies.

Benefits of Pupillometry for Developmental Psychology: In general

There are several advantages to using pupillometry with infants over existing methods. First, pupillometry allows for recording of the same physiological response across different ages in the same tasks. There are numerous studies using pupil size to measure a variety of cognitive processes in adults. If we use similar tasks across populations, we can uncover the development of these cognitive processes. It is important to point out that while maximal pupil size does change with age (Winn, Whitaker, Elliott & Phillips, 1994), most studies tend to calculate change from baseline, which minimizes developmental (or individual) differences in raw pupil size (for more information on proper pupil size preprocessing steps, refer to Kret & Sjak-Shie, 2018). While, in principle, comparison across ages is possible with other dependent measures (e.g., with eye movements, neural responses as measured in event related potential, ERP, and functional near-infrared spectroscopy, fNIRS), in practice, this is very difficult to do as the types

of paradigms that elicit strong and reliable responses in infants and in adults differ. However, evidence of strong and reliable pupillometry responses across disparate developmental ages (e.g., adults and infants) using the same paradigm (Zhang et al., 2018) suggest that in practice, as well as in principle, pupillometry might be more successful in this regard.

Second, pupillometry has the ability to measure the time-course of response within a single trial compared to popular measures in developmental psychology such as looking time and eye movements that are often collapsed or summarized over a period of time. Specifically, eye movements are often analyzed using summary statistics. For example, they are classified as on or off an area of interest or according to the duration of a look during a specific time window. Thus, even though eye movements are sampled with the same temporal resolution as pupillometry, the dependent measures used during a trial do not often reflect the time-course of processing (for a more extensive review of the challenges with measuring looking time, see Aslin, 2007) (see **Figure 1** for an example of how looking time and pupillometry can be used in the same study to display complimentary information). Pupillometry, on the other hand, is meaningful at each time point, which allows you to exploit the high temporal resolution of eye tracking to **examine how cognitive processes unfold over time**. In this way, pupillometry can be used similarly to ERP without many of the constraints of ERP. Furthermore, pupillometry can also be used in conjunction with other measures such as looking time and eye movements. Thus, incorporating pupillometry as a method in developmental psychology research will complement existing methods and extend mechanistic investigations of infant cognition. We specifically propose that pupillometry will greatly facilitate the investigation of prediction in infancy.

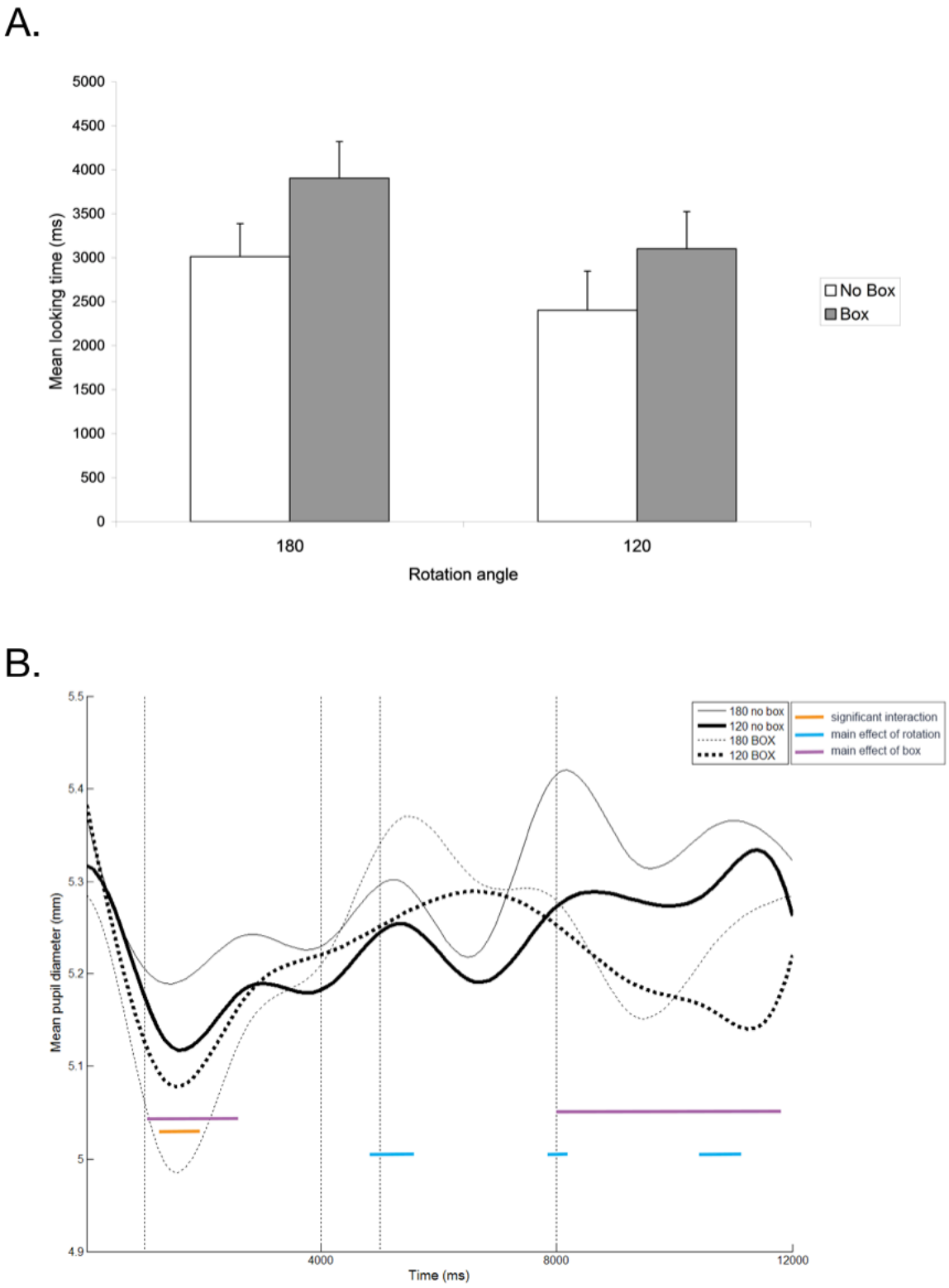


Figure 1. Example of how mean looking time and pupil diameter can be used in the same study. In this object permanence study, 10-month-old infants watched videos of a

drawbridge rotating to either 180° or 120°, either with or without the presence of a box in its path. The 180° rotation is the impossible event because the drawbridge should be blocked by the box and not be able to pass. The 120° rotation event is not impossible because the drawbridge rests on top of the box **(A)** The mean looking time results. Analysis revealed a significant main effect of angle such that infants looked longer at the 180° events. **(B)** The pupillometry results. Dotted vertical lines indicate key moments for 180° conditions (in order: panel starts rotating away from infant, panel stops rotation, panel starts rotating toward the infant, and panel stops rotating). In 120° events, the panel stops its first rotation 1 sec earlier (at 3,000 ms) and begins its second rotation 1 sec later at 6,000 ms. The pupillometry results allowed the authors to pinpoint exactly when similarities and differences occur in the events. Figures are reprinted with permission and modified from Sirois & Jackson (2012).

What is Prediction

Prediction is a cognitive process linking past experience to the interpretation and response to current sensory input. Prediction encompasses several different components (Bubic, von Cramon, & Schubotz, 2010). First, **anticipation** is the process of formulating an expectation that will facilitate the processing of incoming information (LaBerge, 1995). Specifically, an **expectation** is a representation of what you think will happen in the short-term (i.e., the near future), and this representation is sent to sensory or motor areas in the brain. Second, your expectation and the sensory information you actually receive from the real world is compared. If any mismatch between your expectation and sensory information occurs, a **prediction error** is formed (Friston, 2005; Hollerman & Schultz, 1998; McClure, Berns, & Montague, 2003; Rao & Ballard, 1999). Third, prediction errors can be the spark in attention and learning that change expectations to be more accurate in the future; this process is known as **updating**. For the purpose of this review, the term "prediction" refers to future thinking about imminent events, and anticipation, prediction error, and updating refer to the subcomponents of this general future thinking process.

Predictive Processing in Adults

Predictive coding theory, a prominent framework explaining how and why prediction is used in perception (Friston, 2005; Rao & Ballard, 1999), states that the adult brain is extremely sensitive to prediction error such that only the prediction error signal, instead of the predictable input, is transmitted to other areas of the brain. This filtering of information minimizes redundancy and allows for the brain to function efficiently in an ever-changing environment. This efficiency also manifests itself in behavior by resulting in more efficient processing of sensory information (Bar, 2007) (e.g. quickly interpret ambiguous information based on past experiences), which can improve performance (e.g. faster response) (LaBerge, 1995). We will now review studies that have used pupillometry to study prediction in adults (for general reviews of prediction in adults, refer to Bubic et al., 2010; de Lange, Heilbron, & Kok, 2018; Summerfield & de Lange, 2014; Summerfield & Egner, 2009).

Anticipation and Expectation

To the best of our knowledge, there are no adult pupillometry studies that investigate anticipation and expectation, likely due to methodological challenges. Specifically, it is difficult to isolate anticipation from the cognitive load of anticipation. Higher cognitive load has been shown to result in larger pupil change (Ahern & Beatty, 1979; Alnæs et al., 2014; Borghini & Hazan, 2018; Hyönä, Tommola, & Alaja, 1995), and the more an individual anticipates, the more likely they are to engage in increasing cognitive load. Furthermore, pupil change is also modulated by uncertainty (Lavín, San Martín, & Rosales Jubal, 2014; Preuschoff, Hart, & Wolfgang, 2011). Since the manipulation of certainty is one of the principle methods by which

expectations are elicited in infants, it is difficult to dissociate pupillometry responses in anticipation of a stimulus from stimulus uncertainty and cognitive load.

Focusing on expectation, it is difficult to investigate sensory expectations using pupillometry because pupil size is not a measure that can be used to qualitatively differentiate between, for example, a dog's bark versus a cat's meow, independent of other tasks demands. Until it is demonstrated that there is a way to dissociate anticipation from cognitive load and uncertainty, we propose that the best possible alternative to investigating expectation is the use of imaging techniques such as fNIRS and fMRI that can measure patterns of activity in perceptual or motor systems. To sum up, no adult pupillometry studies have specifically investigated anticipation and expectation. It is not clear how to use pupillometry to measure these components of predictive processing, given that it is difficult to disentangle them from cognitive load and uncertainty.

Prediction Error

By contrast, several studies have found a link between larger pupil change and prediction error in adults. Preuschoff et al. (2011) had participants complete an auditory gambling task and measured pupil change during distinct stages of the task, to disentangle decision making, uncertainty and reward from each other. Preuschoff et al. (2011) found that pupil change signaled errors in judging uncertainty (i.e. risk prediction error). Specifically, pupil change was largest when participants lost on low risk (i.e. high chance of winning) trials and when participants won on high risk (i.e. low chance of winning) trials. In other words, **pupil change was largest when the participants' expectation did not match the outcome** when considering both outcome and context (i.e., certainty or risk). Similarly, Braem, Coenen,

Bombeke, van Bochove, and Notebaert (2015) measured participants' pupil size while they completed a flanker task. The results revealed larger pupil change following difficult correct trials than easy correct trials and larger pupil change for easy incorrect trials than for difficult incorrect trials. This study further corroborates the idea that larger pupil change reflects a mismatch between a participants' expectations of an outcome and the actual outcome rather than the outcome per se (e.g., errors produce stronger pupil responses over correct responses). Moreover, pupil change reflects different magnitudes of prediction error. In a predictive-inference task in which participants had to guess the upcoming number, Nassar et al. (2012) found that the larger the error magnitude between the estimate and actual number, the larger the pupil change. This study suggests that pupil change is an extremely sensitive measure of prediction error or the mismatch between expectations and current input.

Pupil change also reflects prediction error in paradigms that do not require explicit responses. For example, Scheepers, Mohr, Fischer, and Roberts (2013) used pupil change to investigate how different types of limerick violations affected a reader. Participants listened to four limericks while having their pupil sizes recorded. The researchers manipulated semantic, syntactic, rhyme or metric violation and found that only the rhyme violation led to a larger pupil change, despite the fact that all violations were reliably detectable. This study suggests that listeners have the strongest expectation for limericks to rhyme and experienced the strongest prediction error when the expectation was violated.

Lastly, pupil change has been linked to the Pearce-Hall theory, an error driven learning theory (Pearce & Hall, 1980; Pearce, Kaye, & Hall, 1982; Pearce & Mackintosh, 2010). The Pearce-Hall theory states that the amount of attention given to a cue, which subsequently improves learning for the cue, is determined by prediction error. Furthermore, after learning the

association between a cue and an outcome, uncertain cues that elicit more prediction error will attract more attention than the consistent cues. In a series of three experiments by Koenig et al. (2018), participants learned to associate different cues with different outcomes. All experiments were designed to examine pupil change during processing of cues that elicited no prediction error (i.e. cues that were consistently followed by the outcome) versus cues that elicited prediction error (i.e. cues that were sometimes followed by the outcome). The researchers were specifically interested in pupil change during anticipation of the outcome and how pupil change during that cue-outcome period changed over the course of learning. Although it is possible to argue that this period is more indicative of anticipation, the researchers noted that pupil change during this cue-outcome period changes as a function of error and, therefore, is more reflective of error instead of anticipation. Overall, Koenig et al. (2018) found evidence that pupillometry fit with the predictions of the Pearce-Hall model: During early learning trials, no prediction error cues caused larger pupil change than prediction error cues, but this effect reversed such that towards the end of the experiment, the prediction error cues caused a larger pupil dilation.

The pattern of pupil change over the course of learning in Koenig et al. (2018) also has overlapping features with the Hunter-Ames model (Hunter and Ames, 1988), a prominent learning model in developmental research. The Hunter-Ames model states that infants exhibit a familiarity preference during the early stages of learning that changes to a novelty preference over time. Tying the two together, adults and infants appear to have an initial preference for the more familiar/consistent event that, over time, turns to a preference for a novel/surprising event. These patterns of results suggest that during a novel experience (e.g. an experiment), humans prefer familiarity or consistency because the opposite is too overwhelming. Through the course of learning, as humans become accustomed to the experience, they prefer the more novel or

surprising event because it provides more information gain. Overall, there seems to be a parallel between infants' familiarity and novelty response with adults' preference for consistent then inconsistent cues, which could be explored with pupillometry.

Taken together, pupillometry can measure prediction error more directly than other dependent measures and has been successfully used in adults, as reviewed in this section. Research studies using a variety of tasks all reach the same conclusion: larger pupil change indicates prediction error or greater mismatch between expectations and outcome. Furthermore, this section illustrates how pupillometry can be used in conjunction with behavioral measures and learning models to investigate how prediction error supports learning in a more systematic way.

Update

To date, few studies have investigated the process of updating via pupillometry, likely due to prediction error and updating being strongly correlated in most experimental designs. Using an extremely clever design, O'Reilly et al. (2013) were able to disentangle surprise and update to examine changes in pupil size as a result of updating. In this study, participants completed a saccadic eye movement task where the goal was to look at the target as quickly as possible and then return to center fixation. Trials were presented in runs such that all the targets appeared in close proximity to each other in a single run. The start of a new run was indicated by changing the color of the target. These trials were used to measure surprise and update because the location of the new target was unpredictable (i.e. surprising) and signalled to participants that all the upcoming targets would show up roughly around the new location. These trials were compared against one-off trials where the target appeared in a random location 25% of the time

and was colored grey. Participants received explicit instructions that one-off trials had no importance for future trials and acted as a pure measure of surprise. O'Reilly et al. (2013) found that pupil dilation showed opposite patterns at different time points for surprise and updating. Specifically, surprise led to a positive pupil change starting at 600ms after target, which aligns with other findings reviewed in the above section. Interestingly, updating led to a decrease in pupil change (i.e. pupil constriction) 1000ms after target. Overall, O'Reilly et al. (2013) showed that it is possible to disentangle surprise and updating with an innovative paradigm and a suitable population.

A couple of studies have used pupillometry to investigate updating, albeit in a more indirect way. While measuring participants' pupil size, Murphy, Van Moort, and Nieuwenhuis (2016) had participants complete an auditory, four-choice reaction time (RT) task where they had to press a letter on the keyboard with a specific finger. For example, every time the participants heard the letter "A," they had to press it with their left middle finger. The study showed that participants had larger pupil changes during trials in which they made a mistake (i.e. experienced prediction error) and, importantly, the magnitude of pupil change predicted RT slowing and improved accuracy on the subsequent trial. In a separate study, Browning, Behrens, Jocham, O'Reilly, and Bishop (2015) investigated how anxiety influenced decision making. Using pupillometry, Browning et al. (2015) showed that different levels of anxious participants did not differ on their response to surprising adverse outcomes but did differ on their response to volatile versus stable environments. Specifically, highly anxious participants were insensitive to whether an environment was volatile or stable and, therefore, had difficulties updating their outcome based on their current environment. These studies suggest that error is adaptive and alters

behavior and, conversely, that a deficit in error detection and updating can also be revealed using pupillometry.

Summary

Overall, prediction is well-studied in adults, and pupillometry is most suited for investigating prediction error. A large number of studies have shown that **prediction error** is related to larger pupil change. Investigations of anticipation and expectation are best investigated using other methods until anticipation can be dissociated from cognitive load. Investigations of updating will require more creative paradigms to dissociate updating from prediction error.

Predictive Processing In Infants

Given the potentially central role of prediction in adult cognition, it is important to determine what role, if any, prediction plays in infant cognition. A number of studies provide preliminary evidence for the same components of prediction in infants as adults (anticipation/expectation, prediction error, updating). Here, we review some of the studies that have investigated prediction in infancy. We have included both studies using pupillometry (mostly in the study of prediction error) as well as studies using methods other than pupillometry.

Studies using eye movements as a dependent measure have established that infants **anticipate** future events. Beginning with classic Visual Expectation Paradigms (VExP) created by Marshall Haith and colleagues in the 1980s, researchers found that young infants can anticipate where images will show up next when the spatial location follows a simple pattern (Canfield & Haith, 1991; Haith, Hazan, & Goodman, 1988). Specifically, infants made

anticipatory eye movements (AEMs) to the location of the target before the target even appeared. The discovery of AEMs in young infants suggest that they are capable of engaging in predictive processing (for a review, see Wentworth, 2009). Recent studies have used AEMs in a more sophisticated manner. For example, Romberg & Saffran (2013) used AEMs to look at how 12 month olds learned from a target that always appeared in the same location versus a target that could appear in multiple locations. They found that AEMs reflected an infants' prior experience. Furthermore, Reuter, Emberson, Romberg, & Lew-Williams (2018) measured AEMs in 12- to 24-month-old infants and found that the ability to update AEMs was correlated with vocabulary size. These recent studies provide a more detailed description of how AEMs actually work and impact infant cognition. Taken together, infants' ability to anticipate is well established in the field.

To the best of our knowledge, Emberson, Richards, & Aslin (2015) is the only study that provides evidence that **expectation,** or a representation of what will occur in the future, can impact the visual system in infants. In this study, Emberson et al. (2015) used fNIRS, an infant-friendly optical imaging technique that uses infrared-light to record changes in the hemodynamics of the brain to investigate top-down sensory prediction in 6-month-old infants as they completed an audiovisual omission task. During familiarization, infants learned to associate audiovisual pairs such that after hearing the auditory cue, they should anticipate the upcoming visual stimulus. After familiarization, unexpected visual omission trials in which the auditory cue was followed by a blank screen were introduced. Results showed that 6-month-old infants had significant visual system activity even during the visual omission trials, suggesting infants formed a visual expectation based on prior experience of seeing a visual stimulus appear after the sound. This provides some evidence of the specificity of the expectations of infants and that they

are not simply anticipating a spatial location but, likely, more specific sensory aspects of their upcoming sensory input.

Prediction error is the most straightforward subcomponent of predictive processing to study in infants as a result of constraints in task design (i.e. must design a task that is engaging for infants) and interpretations of results. Electroencephalography (EEG) studies have found evidence of prediction error in infants. Kouider et al. (2015) designed an audiovisual cueing task where 12-month-old infants learned to associate an auditory cue (sound A, B) with a visual category (face or flower) during familiarization. For example, every time infants heard sound A, they would see a face at the same time (valid trial). After familiarization, infants were introduced to invalid trials 25% of the time where the sound was presented with the other, unassociated visual category. For example, sound A was now presented with a flower. Results revealed an increased ERP response for the invalid trials during the late processing stages. The authors attributed this response to a conscious prediction error signal in infants. Similarly, Ylinen, Bosseler, Junttila, and Huotilainen (2016) found that 12 month olds' ERP responses to novel word forms (interpreted as a prediction error response) correlated with their vocabulary score. These two studies provide initial evidence of neural prediction error early in life.

Similar to adults, prediction error is also where pupillometry is most involved. In a series of studies, Gredebäck and Melinder (2010) first demonstrated that 12-months-old infants who have a good amount of experience being fed had larger pupil change when a feeding action was violated, specifically when the infant saw food reaching a person's hand instead of their mouth. The same action violation did not result in larger pupil change for 6 month olds, suggesting that, as a group, 6 month olds do not experience prediction error. As a follow up, Gredebäck, Lindskog, Juvrud, Green, and Marciszko (2018) approached the same experiment

from an individual differences perspective. The researchers showed 6-months-old infants videos of a person eating and found that some 6 month olds indeed can predict another person's actions by looking at a person's mouth before the spoon reaches their mouth. Interestingly, the infants that made these correct anticipations had larger pupil change when they later saw a violation of feeding actions.

Infants also experience prediction error as a result of unexpected communicative cues. Research has shown that 12-months-old infants understand the goal of pointing is to indicate a referent, which then leads the infant to follow the point and search for the referent (Bertenthal, Boyer & Harding, 2014; Leung & Rheingold, 1981; Morissette, Ricard, & Décarie, 1995). In a recent study by Pätzold and Liszkowski (2019), 12-months-old infants were found to have larger pupil change when they saw a person point to an occluder that was lifted to reveal an empty surface. This omission of a referent violates infants' expectation that one points to a referent. 8 month olds who have been shown to not produce the pointing gesture do not show this effect. These results provide strong evidence for infants having an internal model of certain behaviors (in this case, feeding and pointing) as well as prediction error when their expectations are violated.

Infants' pupil change is also graded and scales with prediction error. Tamási et al. (2016) showed 30-months-old children a picture and an auditory label that matched the picture. The labels were words that 30 month olds knew. The researchers manipulated the auditory label such that there were different levels of label violation. The findings showed that there was an overall pupil difference between correct and incorrect labels, with larger pupil change for incorrect labels. Furthermore, the incorrect labels that deviated more from the correct label had

larger pupil change than incorrect labels that deviated less. These results suggest that, similar to adults, pupil change is also a sensitive measure of prediction error for the younger population.

Infant pupillometry data can also be used in conjunction with error-driven learning models to provide a deeper understanding of infant learning and compare prediction across the lifespan. Zhang et al. (2018) had adults and 6-months-old infants complete an identical implicit learning task designed to help participants learn associations between sounds and pictures. The researchers found significantly larger pupil change for visual omission trials (i.e. trials elicited prediction error) compared to visual present trials (i.e. trials that confirmed participants' predictions) in both age groups (**Figure 2**). Furthermore, the researchers used the Rescorla-Wagner model (Rescorla & Wagner, 1972), another error-driven learning model, and found that the two age groups exhibited similar learning trajectories, suggesting similar predictive processing ability (**Figure 3**). The ability to use pupillometry to make a direct comparison between infants and adults without modifying the experimental task is powerful. We believe pupillometry is the optimal method for researchers interested in examining the development of any domain across different age groups.

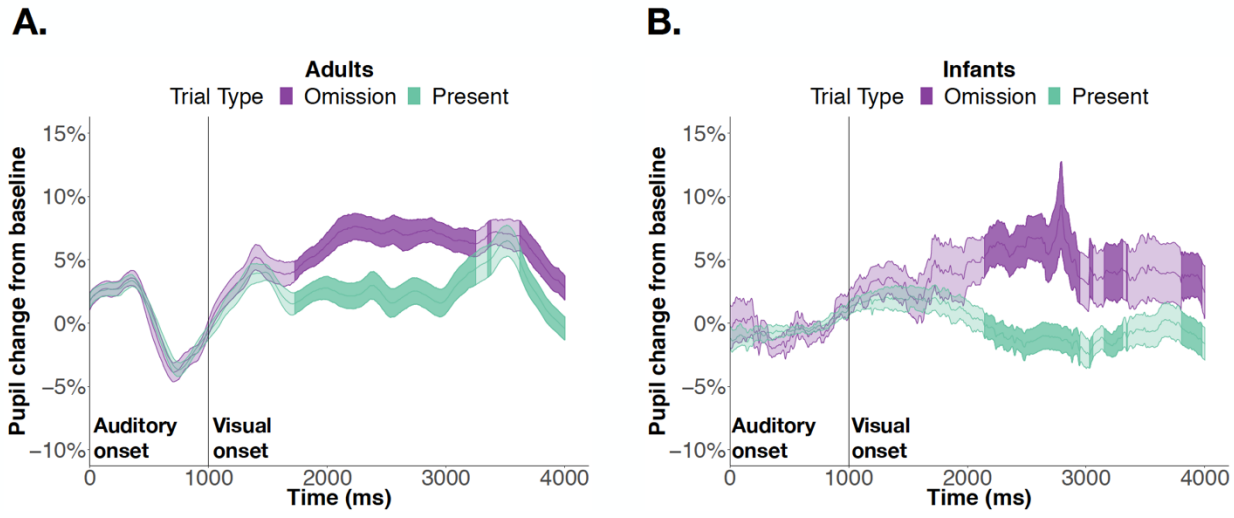


Figure 2. Pupil change for visual omission and visual present trials. (A) Adults **(B)** Infants. Adults and infants completed identical implicit learning tasks. Pupil change was larger for visual omission trials (i.e. prediction error trials) for both age groups. Reprinted with permission from Zhang et al. (2018)

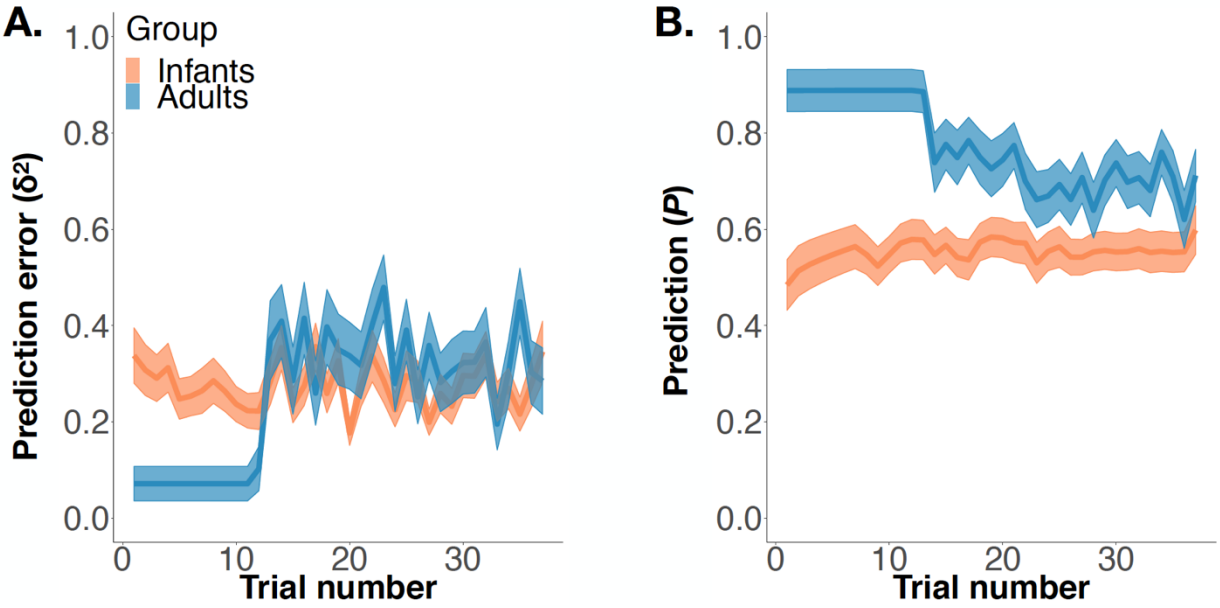


Figure 3. Comparing simulated parameters generated from the Rescorla-Wagner model for adults and infants on a trial-by-trial basis. (A) Prediction error. **(B)** Prediction. Although infants and adults start off distinct from one another, their prediction error and prediction quickly converge. Importantly, the predictions converged accurately (i.e., in this task, the visual stimulus showed up 66% of the time). Reprinted with permission from Zhang et al. (2018).

Related to prediction error is an area of developmental research that uses the violation of expectation (VOE) paradigm. While these concepts are highly similar and may be related, it is not clear whether infants VOE indexes prediction error. The drawbridge study designed by Renee Baillargeon to test infants' knowledge of physical objects (Baillargeon, 1987) is a canonical VOE paradigm. In this paradigm, infants are familiarised to a "possible" event; a drawbridge is then lowered until it is blocked by a box. After familiarization, infants are shown the "possible" event and an "impossible" event in which the drawbridge appears to pass through the box. Baillargeon (1987) found that infants look longer at the impossible event, suggesting

that infants were surprised by the impossible event because it violated the underlying assumptions they had about solid objects. While VOE paradigms have been adapted in many different ways to test infants' knowledge (e.g., from a similar literature: Hespos & Baillargeon, 2006; Ruffman, Slade, & Redman, 2005; Walden, Kim, McCoy & Karrass, 2007; Wang, Baillargeon, & Brueckner, 2004), VOE has not been studied in the way we've framed prediction error. As a result, we are not including VOE experiments as evidence of prediction error in the current paper.

Evidence for young children's ability to **update** comes from Stahl and Feigenson (2015). In this study, researchers found that 11-month-old infants are better at learning sound-object mapping if the object was presented in a situation that was vastly different from their everyday lives (e.g., hovering car) compared to when the object behaved normally (e.g., not hovering). Building upon this study, Stahl and Feigenson (2017) found that 3- to 6-year-old children learned novel words better after experiencing an event that violated their expectations compared to an event that did not violate expectations. Furthermore, the learning was specific, such that it did not transfer to unrelated objects. Both studies conclude that violations of any expectations provide special opportunities for learning.

Summary

Overall, a number of studies have provided preliminary evidence that the same subcomponents of prediction are present in infants and toddlers. However, the majority of these subcomponents remain under investigated. In particular, there are very few direct investigations of expectations (i.e., the content of anticipatory processes), and prediction error and updating are not as amenable to standard developmental methods (e.g., looking time or eye movements) and,

thus, are more rarely studied. The exception here is anticipation: since anticipation can be readily indexed using eye movements, there have been numerous studies demonstrating anticipation in young infants. We will argue in the following section that pupillometry is well-suited to investigate prediction error in infants, and, thus, will substantially expand our knowledge of this currently understudied but very important subcomponent in prediction.

Using Pupillometry To Investigate Prediction In Infancy

By Targeting Different Subcomponents of Prediction

Above, we outlined a model that considers prediction to be comprised of three major subcomponents: anticipation, prediction error and updating. We argue that pupillometry is one of the methods that will allow infancy researchers to investigate these subcomponents of predictive processing in a targeted fashion. Note: All of the studies discussed used phasic pupil change or stimulus-induced increase in pupil diameter relative to a pre-stimulus baseline period.

We also outlined evidence that pupillometry is readily useable to study **prediction error** in infants. There is a general consensus that prediction error is related to larger pupil change for infants and adults, making pupillometry an ideal candidate for researchers interested in examining how prediction error changes across the lifespan. Moreover, pupil change also scales with prediction error, which makes it an excellent candidate for combining with learning models. Error driven learning models allow researchers to quantify and assess prediction error more precisely, which is difficult to do in developmental research due to constraints in task design and duration. Overall, numerous studies using pupillometry to investigate prediction error are summarized above and demonstrate strong potential for this combination of research method and question.

It is also possible to use pupillometry to investigate **updating**, a second subcomponent of prediction. In adults, the use of a highly innovative paradigm with complex task instructions allowed O'Reilly et al. (2013) to investigate updating independent of prediction error in adults. We believe this paradigm might be applied to young children and, potentially, infants though the complexity of the task structure would make this difficult. Another approach mentioned earlier in the review, is to use pupillometry to study updating indirectly. This approach will likely involve comparison of conditions without explicit instructions (e.g. a condition where infants update through exposure of rewarding and highly associated stimuli and another where infants receive random information that makes it impossible to update) and how pupil change differs during these two conditions. There is much to discover about how infants update their internal models. Pupillometry, with its high temporal resolution and ease of comparison across the life span, is likely to be a useful tool though innovative paradigms must be developed to disentangle updating from prediction error.

In both adults and infants, however, there are issues with dissociating **anticipation** from uncertainty and cognitive load (both of which have been shown to modulate pupil size). Moreover, pupillometry cannot be used to investigate **expectations**, as it is not able to distinguish different patterns of activity in particular sensory or motor regions (e.g., pupillometry cannot dissociate visual system responses to a cat and a dog). Thus, until there is a method of measuring anticipation independent of these two other processes, we do not recommend using pupillometry to measure anticipation.

In the Developmental Breadth

Another major contribution of pupillometry to the investigation of prediction is through direct comparisons between different age groups. As developmental psychologists, we are interested in how cognition develops; therefore, comparing the same cognitive process across different age groups is essential. While this is true of most of our methods in principle, in practice, we often have to change paradigms or add tasks to elicit similar behaviors in adults as infants. However, pupillometry appears to be an excellent method to compare infants and adults because the pupil response appears to be largely the same across age groups and similarly automatic, making it an efficient tool for both longitudinal and cross-sectional studies. One consideration to keep in mind is that pupil change may be slower for certain age groups compared to others. Based off of Zhang et al. (2018), the optimal time window to compare prediction error response between infants and adults is soon after the presentation of the stimulus, with the infants' response being slightly delayed (~300ms) compared to adults. Future studies interested in this comparison will have to be especially careful when determining the length of a trial, making sure it is long enough for the pupil response to evolve for all participants.

Sensitivity of Measurement

Pupil change has also been established as a measure sensitive enough to represent **varying degrees of prediction error** in both infants and adults (e.g., the degree of mismatch between an expectation and an outcome or the magnitude of a prediction error). This sensitivity allows pupillometry to advance research interested in finding the optimal amount of prediction

error for learning and in asking more specific questions related to prediction error rather than quantifying it as an all-or-nothing response.

Moreover, the **temporal sensitivity** of pupil change allows it to map how prediction error changes over time. This temporal sensitivity manifests at two different timescales. First, pupillometry can be used to investigate changes over the course of an experiment. In addition to the time-course findings of Zhang et al (2018, **Figure 2**), Koenig et al. (2018) found that, in early trials, the predictable cue elicited larger pupil change while the less predictable cue elicited larger pupil change in later trials. Of course, looking at time-course effects over trials is in principle possible using some other methods available to infancy researchers. This aspect of temporal sensitivity is not unique to pupillometry.

On a smaller timescale, pupillometry is also meaningful at each time point collected within a given trial. This high temporal resolution supports the investigation of how processing unfolds in a single trial, and differs from eye movements, which are sampled from the same resolution but are averaged so each moment is no longer psychologically meaningful. However, pupillometry does have its limitations as it is not well-suited for investigating types of learning or cognition with a strong temporal component (e.g., fast-paced statistical learning of sequential information). Pupil change requires a sufficient amount of time to return to baseline (approximately 6 seconds according to Murphy et al., 2016) to prevent carry over effects. It can, however, be adapted to investigate single violations within a sentence (Byers-heinlein et al., 2017). Overall, pupil change is well suited for measuring how processing changes across different time scales compared to more traditional measures of learning, such as looking time.

Pupillometry Can Be Combined with Other Methods for Convergent Methodological Investigations

Lastly, pupillometry can be readily combined with other dependent measures. Most notably, pupillometry is readily combinable with eye movements (Gredebäck et al., 2018) and computational models (Zhang et al., 2018) to provide convergent investigations into the ways in which the predictability of events affects infants' learning. Employing multiple measures in the same task would provide the unique opportunity to systematically compare such measurements. These comparisons across measures have multiple potential benefits. Given the separate biological underpinning of different eye behaviors (e.g. LC-NE system for pupil size and frontal eye fields for eye movements), similarities or differences between eye-tracking measurements can also reveal more information about neurodevelopment. Moreover, given that different measures might reflect different subcomponents of prediction (e.g., anticipatory eye movements reflect anticipation and pupil response reflects prediction error), we can start to investigate how these subcomponents operate together and develop. With respect to computational modelling, combining it with pupillometry has unveiled novel findings about how infants learn, information we wouldn't have been able to collect without such mathematical models. For example, in Zhang et al. (2018), fitting the RW model to pupil change revealed a learning rate and initial prediction. These model parameters allowed us to quantify how quickly participants learned and what predictions they started the experiment with, respectively, which is difficult, if not nearly impossible, to infer otherwise. Overall, these innovative combinations can promote new ways of framing research questions and expand the field of developmental psychology.

Conclusion

The goal of this article is to provide researchers with an overview of how pupillometry has been used to investigate predictive processing in adults and to better understand how we can do the same in infants. Given the characteristics of pupillometry we discussed above—similar physiological response across the life-span, high temporal resolution and sensitivity, and it being capable of being used in conjunction with other measurements and computational modelling—pupillometry has much to offer developmental research, including innovation. Furthermore, pupillometry can also be used outside of the context of predictive processing (see Goldinger & Papesh, 2012, for a pupillometry review on memory; see van der Wel & van Steenbergen, 2018, for a review on how pupillometry has been used to study cognitive control). Pupillometry truly has the potential to advance the research of prediction and its subcomponents, namely prediction error and other cognitive processes in infants.

Chapter II: Prediction in infants and adults: A pupillometry study[3]

To function efficiently in a changing environment, cognitive systems of adults use past experiences to make predictions about future sensory input. Correct predictions allow efficient processing of incoming stimuli (Kersten & Yuille, 2003; Kutas, DeLong & Smith, 2011). Inaccurate predictions can elicit a prediction error (Hollerman & Schultz, 1998; Rao & Ballard, 1999; Friston, 2005) which promote learning through updating of an internal model (Den Ouden, Friston, Daw, McIntosh & Stephan, 2009; Nassar et al., 2012).

While research has provided neural (Emberson, Richards & Aslin, 2015; Kouider et al., 2015) and behavioral evidence (Canfield & Haith, 1991; Gredebäck, von Hofsten, & Bourdreau, 2002; Lew-Williams & Fernald 2007; Hespos, Gredebäck, von Hofsten, & Spelke, 2009; Romberg & Saffran, 2013) that infants also make predictions, there are reasons to believe infant's predictive capacity is different from an adult's. For example, the infant brain lacks the neural connections found in the adult brain (Nagy, Westerberg & Klingberg, 2004; Smyser et al., 2010), which are used to compare predictions to sensory inputs. Moreover, research on statistical learning, an ability closed related to prediction[4], has shown mixed results when comparing infant and adult performance (differences: Arciuli & Simpson, 2011; Raviv & Arnon, 2017; Emberson, Misyak, Schwade, Christiansen & Goldstein, in revision; no differences: Kirkham, Slemmer & Johnson, 2002; Saffran, Newport, Aslin, Tunick & Barrueco, 1997).

---

[3] A version of this manuscript has been accepted for publication. Zhang, F., Jaffe-Dax, S., Wilson, R. C. & Emberson, L. L. (2018). Prediction in infants and adults: A pupillometry study. *Developmental Science*, e12780.

[4] Statistical learning, an ability related to the extraction of statistical regularities in the environment, is thought to be closely related to prediction both in the purported use of prediction to learn in these tasks and for the ability to use statistically-learned information to generate predictions.

This past work leads to a question: Does an infant's predictive capacity represent an ability that is continuous across the life-span or is it different from an adult's? A discontinuous perspective suggests studies reporting prediction in infancy is fundamentally different than prediction in adults, due to differences in neuroanatomy and knowledge. From a continuous perspective, both infants and adults have highly similar capacities to generate predictions. Continuity suggests that prediction is part of the mechanism we use to build our initial models of the world and not simply update existing models in adulthood.

The goal of this study is to provide the first direct comparison of prediction and prediction error across infants and adults using pupillometry, the measurement of pupil diameter, and identical behavioral paradigms, to begin to answer these broader theoretical questions. The question of how prediction compares across infants and adults has been investigated in other domains. Hespos et al. (2009) compared infant and adult performance during a predictive reaching task; They found that infants and adults performed poorly when the object occluded. While this work suggests similarity of prediction, a limitation of this paper is the use of different paradigms across the age groups (necessary to prevent adults from performing at ceiling). Using the same paradigm for infants and adults and eye movements, Gredebäck et al. (2002) found that both groups made accurate spatial and temporal predictions for an occluded object. Our current experiment allows a more mechanistic comparison of prediction and prediction error through using a novel task, pupillometry, and computational modeling to examine predictive capacities in infants and adults. Importantly, pupillometry allows for the recording of the *same* physiological response in an *identical* behavioral paradigm across these disparate age groups and more closely index prediction error compared to eye movements.

The pupil dilation response (PDR) has been shown to reflect cognitive surprise and prediction error. For example, Nassar et al. (2012) told adults to minimize errors in predicting the next number in a series and revealed that the larger the prediction error, the larger the PDR (see also Raisig, Welke, Hagendorf & van der Meer, 2010; Scheepers, Mohr, Fischer & Roberts, 2013; Braem, Coenen, Bombeke, van Bochove & Notebaert, 2015).

While PDR of infants has been linked to cognitive surprise (Jackson & Sirois, 2009; Gredebäck & Melinder, 2010; Wetzel, Buttlemann, Schieler & Widmann, 2015), no work has directly linked the PDR and prediction or prediction error. For example, using the expectation-of-violation paradigm, Jackson and Sirois (2009) found that 8-month-old and 10-month-old infants have larger PDR for unexpected events that violate object permanence. Another study found similar PDRs in infants and adults when unexpected sounds involuntarily captured attention in an odd-ball task (Wetzel et al., 2015). While odd-ball studies require some aspect of memory for the detection of unusual sounds, this design cannot isolate contributions of predictive processes *per se*. However, predictions can occur without our awareness (Delong, Urbach & Kutas, 2005; Van Berkum, Brown, Zwitserlood, Kooijman & Hagoort, 2005). Thus, we employed a learning task to study the relationship between prediction error and PDR in infants and directly compare infants' and adults' performance.

We modeled the current task after Emberson et al. (2015) which was designed to investigate predictions, specifically those initiated by top-down neural signals. During familiarization, participants learn to associate audiovisual pairs such that after hearing the auditory cue, they expect to see a visual stimulus. After familiarization, we introduced unexpected omission trials where the auditory cue was followed by a blank screen. This violation of participants' visual predictions should elicit a prediction error. Showing a blank screen instead

of a novel visual stimulus allowed us to rule out the possibility that infants are simply experiencing an error signal reflecting the low-level perceptual features of a different visual stimulus, rather than violation of a prediction for upcoming visual input. Using this paradigm, Emberson et al. (2015) reported that 6-month-old infants exhibited significant visual system activity during an unexpected visual omission. In other words, top-down signals from the auditory system to the visual system resulted in a visual prediction. Thus, both the learning task and the type of violations employed were specifically designed to tap into predictions as opposed to other cognitive processes or types of error.

We investigated three questions:

1) Is prediction error in infants linked to larger PDR? Behavioral and neurodevelopmental research suggest early availability of predictive processing (Romberg & Saffran, 2013, Lew-Williams & Fernald, 2007; Emberson et al., 2015; Kouider et al., 2015), however, no study has used PDR to index prediction error in infants. We hypothesize that 6-month-old infants will show larger PDR in response to omission compared to present trials.

2) How does prediction error (measured using PDR) compare between infants and adults? The current study is the first to use the same paradigm and pupillometry to compare prediction error across developmental time.

3) Is the time-course of learning and prediction similar in infants and adults? There's a lack of research focusing on infants' learning trajectory due to methodological constraints. We combined the PDR collected on each trial and computational modeling to examine learning trajectories across age groups. Specifically, we used the Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972), a well-established associative learning

theory that explicitly includes prediction and prediction error, to uncover the trial-by-trial dynamics of learning and prediction.

Methods

Participants

Adults: We collected data from forty-four adults. One participant was excluded due to age (51), eight were excluded due to technical errors and three were excluded due to experimenter error. Our final adult sample included thirty-two healthy adults (M = 21 years; SD = 2.5; twenty-seven females). We did not collect demographics for one participant but they fall within the standard range for the population.

Infants: We collected data from forty-three infants: Four infants were excluded due to insufficient number of good trials (see Appendix A); Two were excluded because they were later identified as preterm; Ten were excluded due to technical errors. Our final infant sample included 27 infants (M = 5.7 months; SD = 0.60 months; ten females). Experiments were approved by the university's Institutional Review Board and informed consent was obtained before beginning the study from the adult participant or legal guardian of the infant. Adults received a course credit or $6 for their participation. Families received $10, a t-shirt and a children's book for their participation.

Stimuli

Two auditory stimuli and two visual stimuli formed two audiovisual pairs (A1V1, A2V2). Participants learned that A1 and A2 (i.e. the predicting sounds, xylophone and trill) predicted V1 and V2 (i.e. the predicted visual stimuli, **Figure 4**), respectively. We had two

additional auditory stimuli (twinkle, doorbell), which were presented simultaneously with the visual stimuli and indicated when the visual stimulus should appear. Baseline videos (5) were presented to give the pupil enough time to fully evolve (**Figure 4**). To prevent luminance-related changes in pupil size, all visual stimuli were made isoluminant to the background color of the experiment and to the experimental room (see Appendix A).

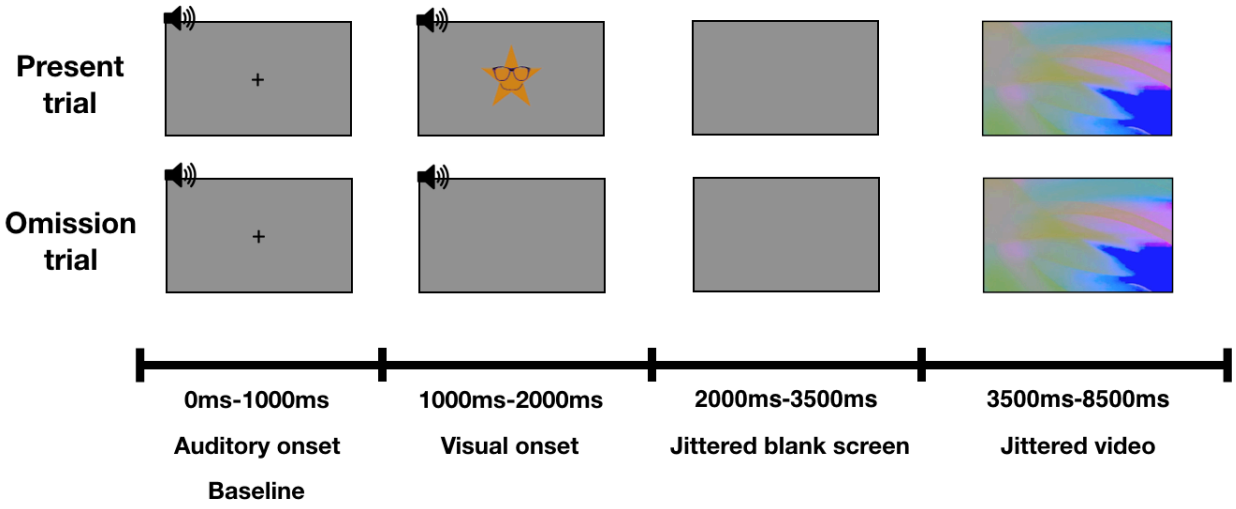


Figure 4. Trial procedure. Arrangement of a present trial and an omission trial. Both trials started with an auditory stimulus that is presented for 1000ms, followed by either a visual stimulus (present trial) or blank screen (omission trial) for another 1000ms, followed by a jittered blank screen and a jittered video.

Design and Procedure

Participants sat approximately 60 cm from the monitor and eye tracker (Eyelink 1000, SR Research, Ontario, Canada, see Appendix A for details on eye tracking procedure). There were two trial types (**Figure 4**): Present and omission. During a present trial, the predicting sound (A1/A2) and fixation cross appeared for 1000ms followed by a visual stimulus for 1000ms that appeared with a coincident sound. The visual stimulus appeared in the center of the screen, and after 500ms slid towards the bottom of the screen and disappeared at 1000ms. Then, participants

saw a blank screen from 900 – 1500ms, designed to capture the pupil response and make an omission clearer. Next, one of five videos played for a jittered interstimulus interval (ISI) of 4.4 – 5s. Each video was shown once before they were shuffled and presented again. The next trial followed the conclusion of video. The arrangement of an omission trial was identical to the present trial, except the predicted visual stimulus was omitted. Stimuli were presented using MATLAB for Mac (R2015b) and Psychtoolbox (3.0.12 Beta, SVN revision 6399).

The experiment began with a familiarization phase of 12 present trials to allow participants to learn the two audiovisual pairs (A1V1, A2V2). After familiarization, participants saw four present trials (two each A1V1, A2V2) and two omission trials (one A1--, one A2--) presented in a randomized order, and then repeated with new random orders for a total of 114 trials. Adults completed all 114 trials, and infants participated until they became consistently fussy or inattentive. This design ensured participants only saw unexpected visual omissions 33% of the time. This percentage balances the need for as many visual omission trials as possible while at the same time allowing participants to maintain association of the two audiovisual pairs. Previous studies examining the visual omission design have used a 20% visual omission rate and found very robust responses with no decline over the course of the experiment (Emberson et al., 2015).

Preprocessing

Details of the preprocessing steps can be found in Appendix A. As a result of preprocessing, we removed 44% or 808 trials for infants and 1.2% or 45 trials for adults. Excluding this number of trials is common for infants. It is preferable to be conservative as trials with blinks, large amount of data missing, etc. can introduce outliers and bias results. For similar

exclusion rate see Fawcett, Wesevich and Gredebäck (2016). We calculated the PDR as percentage change from baseline to accommodate variations in pupil diameter due to individual differences, which is a standard approach for pupillometry data (Preuschoff, Hart & Einhauser, 2011; Lavin, San Martin & Rosales Jubal, 2014; Mill, O'Connor & Dobbins, 2016). We defined baseline as the average pupil diameter during the period after the predicting sound was played and before the visual stimulus appeared, which was 1000ms (**Figure 4**). This baseline measure was subtracted from each pupil diameter measurement and divided by baseline to get the PDR and then multiplied by one hundred to convert the decimal number to percentage.

Modeling PDR Response Using A Learning Model

We used a learning model to examine the trial-by-trial changes of prediction and prediction error in the task. We chose the RW learning model because it measures how prediction error affects the strength of predictions. Therefore, the model will be sensitive to the strength of the predictions our participants have for the visual stimuli, as they view present trials (designed to strengthen predictions) and omission trials (designed to generate prediction error).

We assumed average PDR after visual onset of each trial reflects the magnitude of prediction error or surprise in that trial (Preuschoff et al., 2011; Sirois & Jackson, 2012; O'Reilly et al., 2013; Lavin et al., 2014; Kloosterman et al., 2015). We decided to calculate average PDR rather than taking the change in PDR because it is a less noisy measure, which is important for our younger participants. For each participant, we fitted a generative implementation of a learning model to their average PDR response between visual onset and the end of trial. The model asserts that prediction error is calculated for each trial, t:

$$\delta(t) = O(t) - P(t)$$

where: $O(t)$ is the appearance (or omission) of visual stimulus at trial t and $P(t)$ is the predicted probability of that appearance. In addition, predictions are updated for each trial based on the prediction error and learning rate, where α is the learning rate:

$$P(t+1) = P(t) + \alpha\delta(t)$$

We fitted the individual models by minimizing the Mean Square Error (MSE) between the PDR and δ^2 (level of surprise or magnitude of prediction error regardless of the direction of the error) throughout the experiment for each participant (see Appendix A). Furthermore, fitting the model revealed each participant's learning rate, α, and initial prediction, $P(0)$.

Results

PDR To Unexpected Visual Omissions

We examined the PDR of present vs. omission trials for both age groups using two-tailed t-tests across successive time points and corrected for multiple comparison using a conservative method (the Benjamini & Hochberg, 1995, procedure controlled for the rate of false discovery; Lavin et al., 2014; Mill et al., 2016). We used multiple t-tests to analyze the time-course because we wanted to compare every single time point to identify where the prediction error signal starts and ends. As hypothesized, we found differences in the PDR across trial types for both age groups (**Figure 5**). In adults, the PDR varies across trial type (pFDR < 0.05) for two sustained periods from 1726–3250ms and 3620–4000ms. In infants, PDR vary across trial type (pFDR < 0.05) for a sustained period from 2144-2948ms and two brief periods from 3152-3308ms and 3804-3976ms. Majority of participants from both age groups showed a larger pupil change to omission trials compared to present trials (**Figure 6**) where pupil change is an average calculated from after the visual stimulus appears/supposed to appear (i.e. from the second line of **Figure 5**)

to 4000ms. Overall, these results demonstrate both infants and adults exhibit larger PDR to omission trials compared to present trials with a similar period[5]; Though, infants exhibit a differential PDR ~400ms later than adults.

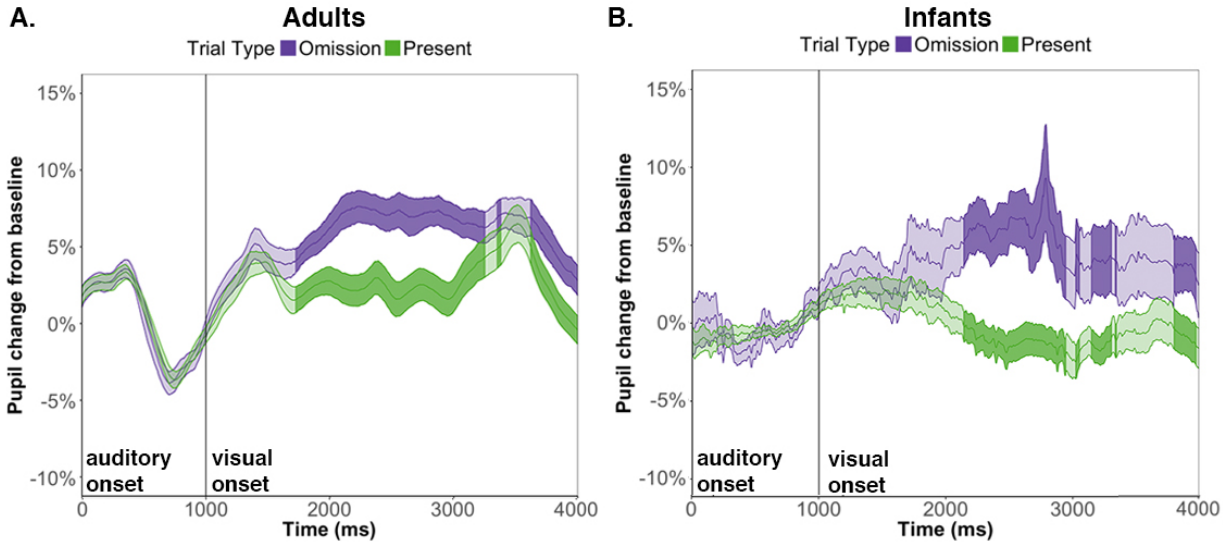


Figure 5. **PDR plotted against time for the two trial types.** Significant results, darkened, are identified after FDR correction, pFDR < 0.05. Data depicted as Mean +/– SEM. The first line indicates on average when the predicting auditory cue plays. The second line indicates on average when the predicted visual stimulus appears/suppose to appear. A. Significant differences in adults were found in three periods: 1726–3250ms, 3350-3378ms and 3620–4000ms. B. Significant differences in infants were found in five periods: 2144-2948ms, 3028-3056ms, 3152-3308ms, 3340-3348ms, and 3804-3976ms.

[5] Prior work has found that the presentation of a visual stimulus can lead to a decreased PDR (e.g., updating response in O'Reilly et al., 2013). However, this is not likely to be driving the difference between the visual present and visual omission trials in the current design. In the current paradigm, after familiarization, the trial type that is most likely to drive an updating response in the omission rather than the present trials since the present trials are highly familiar and most frequent. Furthermore, anecdotally, we do find a negative response upon visual presentation but it is only found in the first trial of the task (the first familiarization trial). This negative response is quite pronounced and never observed for any other trial.

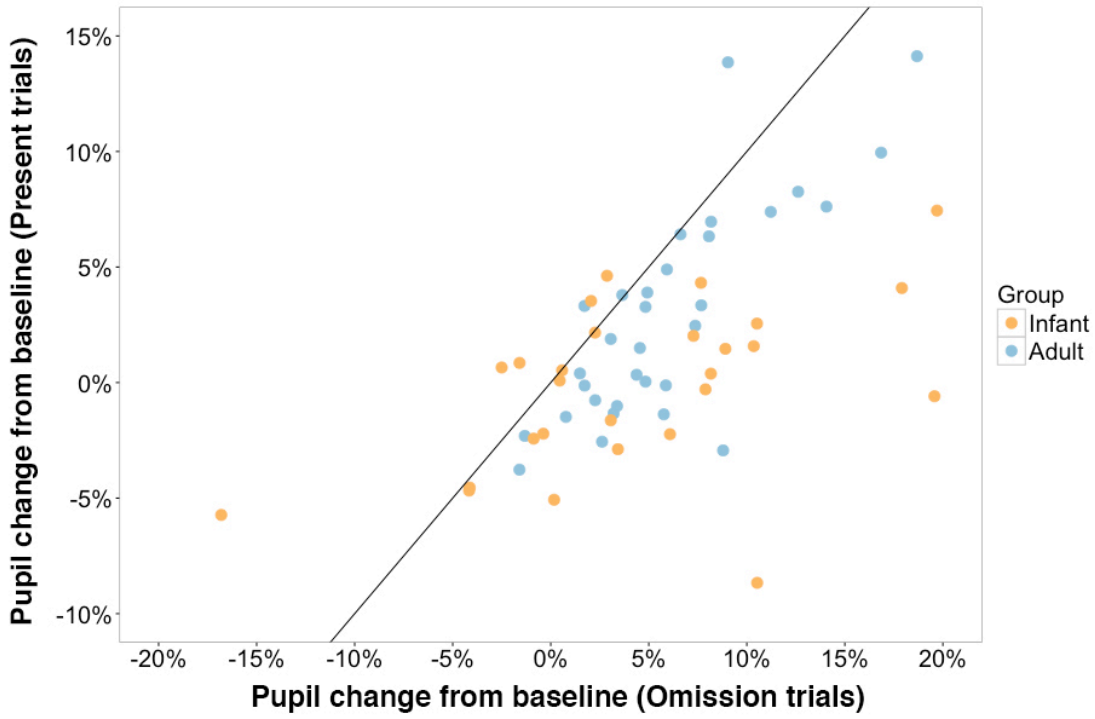


Figure 6. **Scatterplot of average PDR (average calculated from immediately after the visual stimulus appears/supposed to appear to 4000ms) to present trial versus omission trial, for both age groups.** The diagonal line represents equal average PDR to both trials. Participants below the diagonal line show greater PDR to omission trials than present trials. Participants above the diagonal line show greater PDR to present trials than omission trials.

Learning Model

We calculated the average number of trials infants completed (37 trials) and used only the first 37 trials of adult data to make sure the exposure to the experiment was as close as possible for both age groups. Before interpreting the results, we determined model fit by correlating average PDR and prediction error magnitude (δ^2) from the model for every trial of every subject (**Figure 7**). Results revealed significantly positive correlation for both infants and adults (adults: $r = 0.19$, $p < 0.001$; infants: $r = 0.16$, $p < 0.001$), suggesting the larger the prediction error, the larger the PDR. Therefore, PDR is a good index of prediction error.

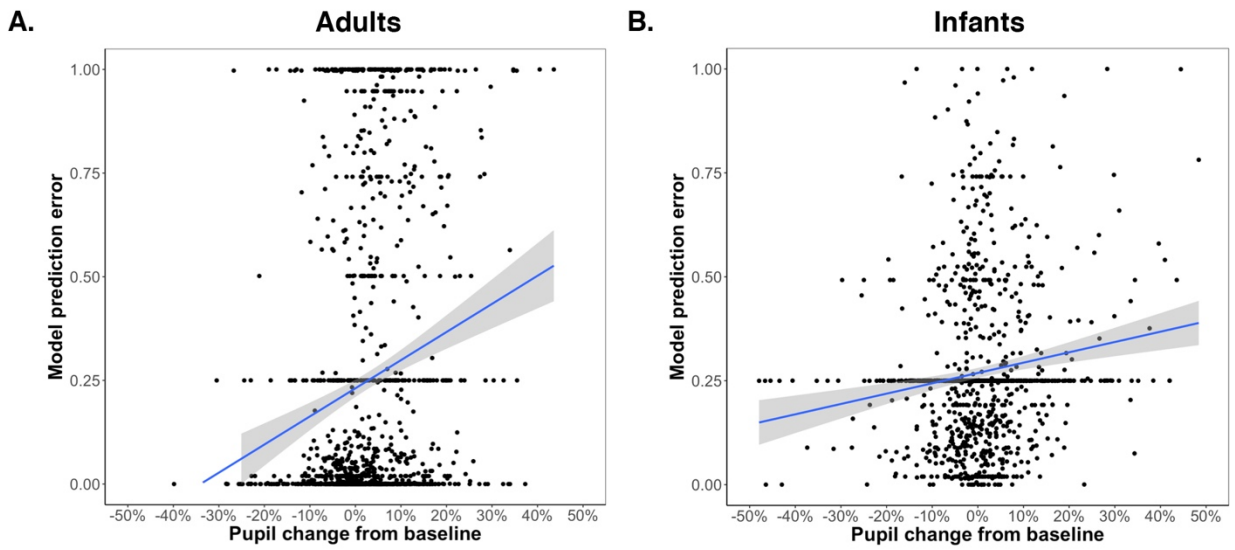


Figure 7. Correlation of average PDR and prediction error magnitude (δ^2) from the model for every trial of every subject to determine model fit. A. Results revealed significantly positive correlation for adults. B. Results revealed significantly positive correlation for infants.

Comparing adults and infants, we found the two groups did not differ on the goodness of fit, as measured by the MSE (adults: 1.44 ± 0.054; infants: 1.41 ± 0.042. U = 413, p = 0.78). Comparing the parameters (**Figure 8**), a Mann-Whitney test indicated adults had significantly higher initial predictions, $P(0)$, of a visual stimulus than infants (adults: 0.88 ± 0.044; infants: 0.48 ± 0.052. U = 107, p < 0.0001). In other words, adults came into the experiment with higher expectation of audio-visual association or the presentation of a visual stimulus than infants. Additionally, adults had significantly higher learning rate, α (adults: 0.31 ± 0.068; infants: M = 0.055 ± 0.032. U = 265.5, p < 0.05. Mean ± SEM. Mann-Whitney tests), which is likely due to adults beginning the experiment with a strong prediction for the presentation of a visual stimulus that they needed a larger learning rate to update their predictions when encountering their first omission, whereas infants do not. However, a comparison of the infant learning rate to zero confirmed that infants are indeed learning (Z = 4.53, p < 0.0001. Wilcoxon signed-rank test).

Once the parameters (learning rate and initial prediction) were estimated, we examined prediction and prediction error on a trial-by-trial basis (**Figure 9**) which revealed that, although infants and adults start off distant from one another, they quickly converge. The divergence at the beginning of the experiment for δ^2 and P can be explained by the significant difference in initial prediction or $P(0)$ between infants and adults. However, as the experiment continued, both groups eventually experienced similar level of surprise and generated similar prediction. Importantly, the predictions converged accurately (i.e., a visual stimulus shows up 66% of the time and their predictions converge on this probability).

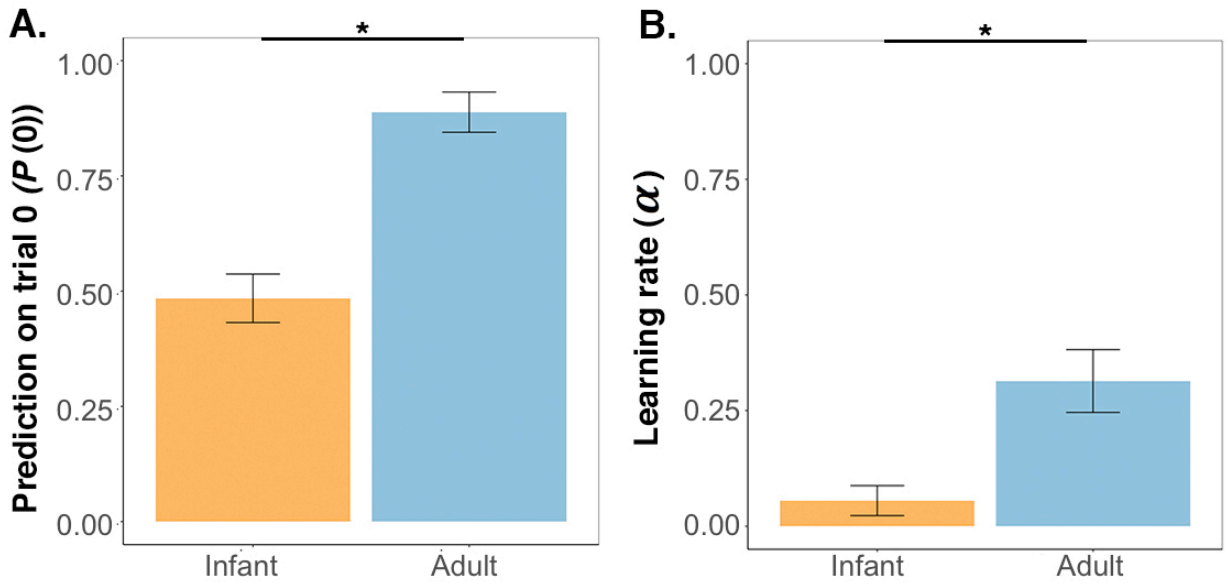


Figure 8. Comparing parameters from the learning model for infants and adults. A. Comparing initial prediction, $P(0)$, between infants and adults. Adults had significantly higher initial predictions than infants. B. Comparing the learning rate, α, between infants and adults. Adults had significantly higher learning rate than infants.

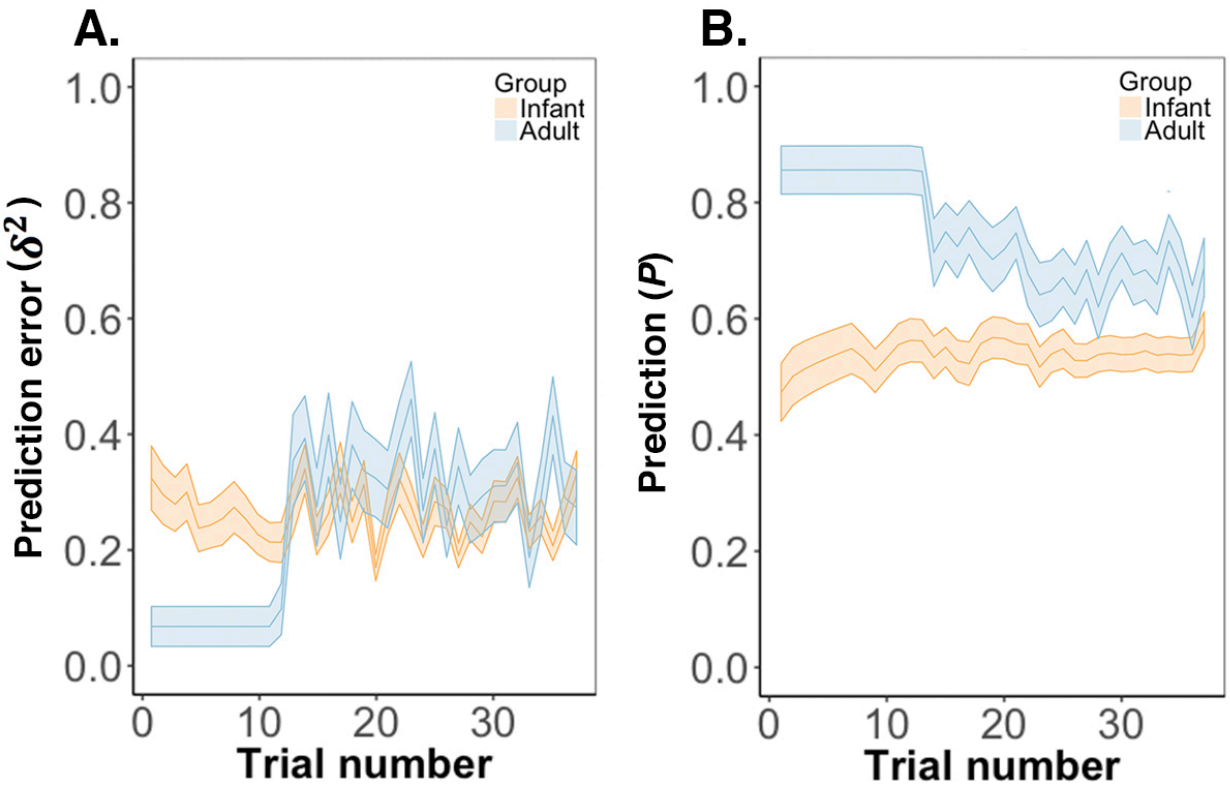


Figure 9. Comparing simulated parameters generated from the infant and adult learning model on a trial-by-trial basis. A. Prediction error. B. Prediction.

Discussion

In this study, we investigated continuity of predictive processing in young infants and adults. While studies have compared prediction across these ages, no study has used pupillometry and computational modelling to examine predictive capacity over developmental time. Consistent with the continuous perspective, we found larger PDR during omission trials compared to present trials in both age groups, with the exception that the divergence of the PDR between trials was slightly delayed in infants (**Figure 5**). Furthermore, both groups generate prediction errors (δ^2) and use it to adjust their prediction (P) to the correct stimulus probability as revealed through modeling trial-by-trial PDRs with the RW model (**Figure 9**). Moreover, results from our task, which has been shown to elicit top-down signals (Emberson et al., 2015),

suggest infant prediction error might be generated from top-down signals, similar to adults. Thus, using the same learning task and recording the same physiological response, we find that prediction abilities in young infants are continuous with the prediction abilities present in adults.

While our results also reveal differences between infants and adults, we believe they do not suggest discontinuity of predictive processing. The modelling results for predictions on trial 1 (and by consequence their learning rates) are different between the two groups. Despite our efforts to create a novel learning context for both infants and adults, we believe the differences are a result of the different experiences infants and adults have about the world (i.e. about sitting in front of a computer screen). Moreover, in **Figure 5**, after the first significant PDR period (adults: 1726-3250ms, infants: 2144-2948ms), the time-course graph has other significant PDR periods (adults: 3350-3378ms and 3620–4000ms, infants: 3028-3056ms, 3152-3308ms, 3340-3348ms, and 3804-3976ms), which are different for infants and adults. These differences are likely due to the protracted PDR, which likely varies within and across age groups. These individual differences will manifest itself statistically as intermittent periods of significant difference. Despite these differences, this doesn't constitute prediction or prediction error being discontinuous across these age groups. The similarities between groups are in the time-course of the PDR, the fit of the RW learning model to the PDR response and the convergence of expectations towards the probability of the visual event. These measures are directly related to prediction and, given their continuity across the two groups, we conclude there is continuity of predictive processing despite differences in initial expectations. However, future research should extend investigation to other age groups as there could be developmental changes before 6 months and a non-linear trajectory between 6 months and adulthood.

As always, it is important to consider whether results could be due to other processes that give rise to a behavior that looks like prediction. Given the clear correspondences between the PDR on each trial and the prediction error parameter of a foundational learning model (RW model), this isn't a parsimonious explanation. Moving forward, we can start to ask more specific questions about the use of prediction to construct initial models of the world. For example, how stable are infants' predictions and models of the world compared to adults?

Our study adds to the growing field of literature examining predictive capacities between infants and adults (Gredebäck et al., 2002; Hespos et al., 2009) as well as the nature of the PDR. We provide the first evidence that the infant PDR can index predictive processes and therefore is not only reactive (see Nassar et al., 2012; O'Reilly et al., 2013 for evidence in adults; Gredebäck & Melinder, 2010; Gredebäck & Daum, 2015 for information about the distinction). This contrasts with studies that have found that the PDR is not reflective of prediction in the context of action perception (Gredebäck & Melinder, 2010; Gredebäck & Daum, 2015). We find that the PDR can provide fruitful information regarding predictive abilities in other domains. Future work should continue to investigate this predictive PDR making sure to dissociate it from processes such as cognitive load and memory retrieval.

In sum, using the same learning task for both adults and infants, we found significantly larger pupil dilation (PDR) during trials that violate predictions compared to trials that met predictions. Specifically, the strong parallels in time-course and magnitude of the PDR, along with modelling results, suggest a continuity of predictive abilities across these ages. Taken together, our data suggests that research documenting prediction in young infants (Lew-Williams & Fernald, 2007; Romberg & Saffran, 2013; Emberson et al., 2015; Kouider et al., 2015) is

uncovering a life-span ability and not simply a nascent form of prediction that isn't yet comparable to adults.

Chapter III: Opposing timing constraints severely limit the use of pupillometry to investigate visual statistical learning [6]

Statistical learning (SL) is the ability to exploit patterns in the environment after passive exposure. SL has been demonstrated using a wide range of stimuli (e.g., Aslin, 2017; Saffran & Kirkham, 2018) and is considered a domain general learning mechanism used by cognitive systems to detect structure in the environment. Since SL is available starting early in life, it is believed to be important for development (Armstrong, Frost, & Christiansen, 2017; Krogh, Vlach, & Johnson, 2013). This paper focuses on visual statistical learning (VSL), or statistical learning paradigms using visual stimuli and investigates whether pupillometry (the measurement of pupil diameter), could be effectively used to investigate the mechanisms supporting VSL.

A standard VSL paradigm is composed of two phases: Familiarization and test. During familiarization, participants are shown a stream of images one at a time, for ~1000ms, without explicit instructions. The images are organized based on statistical information (e.g., transitional probability, co-occurrence frequency) such that some of the images are more likely to appear one after another compared to others. For example, a stream of images can be composed of triplets, such that the same three images always appeared in the same sequence (ABC-DEF-ABC-GHI) and statistical information is the only cue that learners have to uncover the structure or organization of the visual stream. The second phase is test phase, where participants' learning of the statistical information is probed by forced-choice judgments (e.g., A vs B) (Brady & Oliva, 2008; Turk-Browne & Scholl, 2009; Turk -Browne, Jungé, & Scholl, 2005) or familiarity ratings for single items (Baker, Olson, & Behrmann, 2004; Turk-Browne, Scholl, Johnson, & Chun,

---

[6] A version of this manuscript has been accepted for publication. Zhang, F. & Emberson, L. L. (2019). Opposing timing constraints severely limit the use of pupillometry to investigate visual statistical learning. *Frontiers in Psychology*.

2010). Following common nomenclature in the field, we will refer to these measures as **offline measures** because they are collected after learning (i.e., after the familiarization phase or during the test phase) has occurred. Conversely, we will refer to measures collected *during* learning (i.e., during the familiarization phase) as **online measures**.

Majority of the SL research conducted used only offline measures, and these studies have revealed a lot about the extent of SL, but are best suited for investigating what participants have learned rather than how they have learned. For example, VSL can operate over a variety of visual features (e.g., color and shape) and can even bind these features and objects (Turk-Browne, Isola, Scholl, & Treat, 2008). Furthermore, offline measures such as looking time have revealed that newborns (i.e., 1-3 day-old infants) are also able to learn simple visual statistical information (Bulf, Johnson, & Valenza, 2011).

Offline measures, however, have their limitations (for a review see Siegelman, Bogaerts, Christiansen, & Frost, 2017). One of the biggest limitations, which is the focus of this paper, is that it is very difficult to use offline measures to investigate the online learning mechanisms that support SL. Important questions such as "what mechanisms allow for such rapid learning of this statistical information?" and "how do mechanisms that support SL change throughout development?" are still points of active investigation, in part, because of the reliance on offline measures.

While it is likely that multiple cognitive processes such as attention, memory and prediction interact online to enable VSL, this paper focuses on how **prediction**, or the ability to use past experiences to generate expectations about future sensory input, supports statistical learning. This work unifies theories that have proposed that prediction is a core aspect of adult brain function (Clark, 2013; Friston, 2005; Rao & Ballard, 1995) and SL studies that suggest

prediction is at play (Kim, Seitz, Feenstra, & Shams, 2009; Siegelman, Bogaerts, Kronenfeld, & Frost, 2018; Turk-Browne et al., 2005). However, it is very difficult to investigate the role of prediction in VSL using offline measures alone.

The shortcomings of offline measures can be overcome by including online measures into SL paradigms. Online measures lend themselves to investigating how representations are formed during learning. For example, if a measure is able to reflect a process like prediction, using such measures can provide a window into how prediction operates during SL. Furthermore, online measures are helpful for revealing the trajectory of learning, both between trials and on a trial-by-trial basis within an individual, without repeated testing.

Recent SL studies have started using online measures. Electroencephalography (EEG) is a common tool used to record online measures of SL (Abla, Katahira, & Okanoya, 2008; Abla & Okanoya, 2009; Batterink & Paller, 2017). Event-related potentials (ERP) such as the N400 have been shown to be different for good versus poor learners (Abla et al., 2008; Abla & Okanoya, 2009). Furthermore, reaction time (RT) is the most common behavioral online (Baker et al., 2004; Franco, Gaillard, Cleeremans, & Destrebecqz, 2015; Gómez, Bion, & Mehler, 2011).

Online measures can also be combined with offline measures in a single SL task. This combination is a powerful approach to examining the underlying cognitive processes that support learning and how that intertwines with what people learn. Indeed, Siegelman et al. (2018) found that the online measure of RT predicted offline measures of learning: Participants who had faster RTs for predicted stimuli also performed better in the test phase (for similar results using a self-paced artificial grammar learning task see Karuza, Farmer, Alex, Smith, & Jaeger, 2014).

Despite the usefulness and broad use of RT as an online measure of statistical learning, it has substantial drawbacks. First, RT is most commonly gathered through button pressing, and this type of task is only possible for some populations. For example, button pressing cannot be used to investigate SL of infants and is very difficult for young children. Second, it has been argued that pressing buttons during familiarization interferes with learning (Franco et al. 2015; Gómez et al. 2011). Relatedly, Turk -Browne et al. (2005) found that divided attention due to concurrent tasks can lead to poor learning as measured in offline tests. Finally, it has been suggested that RT measures are not stable. For example, Franco et al. (2015) found a different pattern of RT evolution in their click detection task compared to Gómez et al. (2011).

Building from these limitations of RT, an ideal online measure is one that reflects learning but requires no overt action or task and can be used in a wide-range of populations, ideally under identical experimental conditions. To this end, we propose **pupillometry**, or the measurement of pupil diameter, as an online measure for investigating SL. Pupillometry has none of the disadvantages of button pressing listed above and several additional advantages. First, pupillometry does not interfere with the learning process because it does not involve a secondary task. Pupillometry is automatically recorded from an eye tracker at extremely high temporal resolution. Second, due to the high temporal resolution, we record multiple measures during the same trial unlike RT, which is a summary measure. Sampling online processes throughout the presentation of a single stimulus can help to differentiate online learning mechanisms (e.g., processes that occur during stimulus processing versus in the anticipatory periods before a stimulus). Lastly and most importantly, changes in pupil diameter are time-locked to events and reflect processes that are part of prediction such as prediction error (i.e., incorrect predictions, Koenig, Uengoer, & Lachnit, 2018; Nassar et al., 2012; Zhang, Jaffe-Dax,

Wilson, & Emberson, 2018) and updating of predictions (O'Reilly et al., 2013). Since pupillometry has been found to be sensitive to predictions, it may be able to evaluate the hypothesis that prediction is part of the learning mechanisms supporting SL.

The following studies are the first set of experiments to use pupillometry as an online measure during a VSL task. In a standard VSL task, a trial is composed of an image followed by the ISI. In the current study, a trial is composed of the ISI followed by the image. In other words, we are taking the ISI from the previous shape and considering it as the anticipatory period for the current image which is presented during the viewing period of a trial (**Figure 10**). One might ask: How do we know the pupil diameter during the anticipatory period is reflective of anticipation and not processing the visual stimulus from the previous trial? We can simply observe the pupil change to make that distinction. If the anticipatory period reflects lower-level visual processing, we would expect to see identical pupil change for all images. But if pupil change reflects prediction, then we expect to find pupil differences relative to image position in the triplet, and this difference may occur either in the anticipatory period or viewing period.

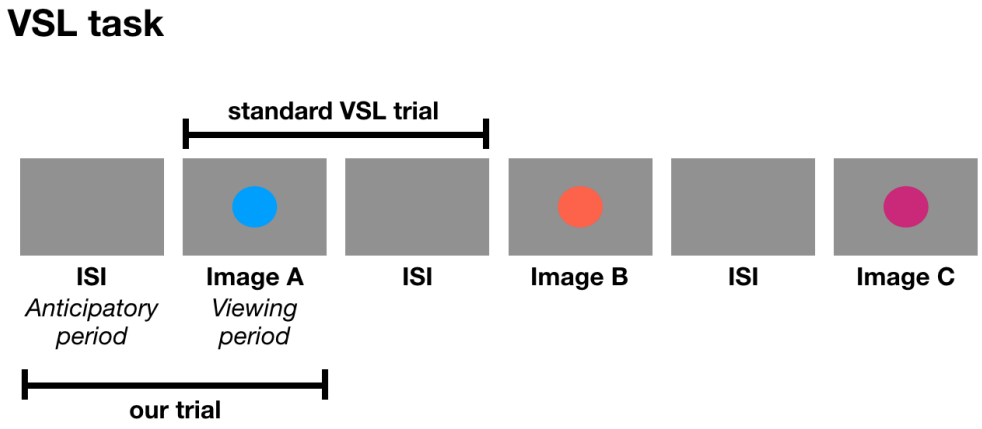


Figure 10. The formation of a trial in a standard VSL study compared to a trial in the current study.

To compare pupil diameter within and across participants, we calculated pupil change from baseline for all experiments. We designated baseline as the anticipatory period of the first image in a structured sequence (i.e., triplets and pairs). This is a standard approach for pupillometry data (Eckstein, Guerra-carrillo, Singley, & Bunge, 2016; Kloosterman et al., 2015; Lavín, San Martín, & Rosales Jubal, 2014; Preuschoff, Hart, & Wolfgang, 2011) and accommodates variations in pupil diameter due to individual differences and the dynamic changes that occur over the experiment.

If pupillometry reflects online mechanisms supporting VSL, we hypothesized that several factors that would influence pupil diameter:

1. **Image position within the triplet:** Based on previous studies demonstrating faster RT for predictable stimulus compared to unpredictable stimuli, and research suggesting that larger pupil diameter reflects prediction error, we hypothesize that the pupil diameter will differ for each image in a triplet during the viewing period. Specifically, we hypothesized that the pupil diameter will be the largest for the first image in the triplet (i.e., the least prediction image/largest prediction error) and smallest for the third image in the triplet (i.e., the most predictable) and the second image to fall in between. In our previous work with pupil size and passive learning in infants and adults, we find greater pupil size during violation trials (similar to unpredictable) compared to standard or predictable trials (Zhang et al., 2018). It is also possible that image position will influence pupil change during the anticipatory period, which would suggest that image position plays a role in temporal prediction (i.e., when will the visual stimulus appear). However, based on our previous work and other pupillometry work, we

hypothesized the effect would show up during the viewing period if participants were predicting which image will appear instead of when.

2. **Number of times seen (i.e., familiarity effect**): During the familiarization phase of SL, participants are exposed to the stimuli multiple times. Research has shown that pupil dilation to be larger for familiar stimuli compared to novel stimuli (Kafkas & Montaldi, 2012, 2015). Therefore, we hypothesize that the more times you view an image, the larger the pupil change during the viewing period.

3. **Relationship to offline measures** (e.g., test accuracy): If pupil diameter is a suitable online measure for SL, it should be diagnostic of participants' performance during test phase (i.e., correlated with offline measures). Kafkas and Montaldi (2011) had participants complete a "remember/know" task to examine the relationship between pupil change and memory strength. Participants incidentally encoded a series of images and were later asked to rate recognized items along a "strength" scale and to call unrecognized items "new." The authors found that pupil constriction at encoding was associated with subsequently stronger memories. With respect to our study, we also expect to find pupil constriction to be related to better accuracy during test phase.

To foreshadow our results, while the combination of pupillometry and VSL has the potential to expand the online measures available to investigate VSL, we propose that the timing necessary to produce VSL is too fast to see the pupil effects that we hypothesized. In a series of three VSL studies, we found that pupil differences between image positions within a triplet to be a result of where we chose to baseline instead of online processes like prediction. Furthermore, we did not find pupil diameter to be related with offline measures of learning. We conclude the

paper by discussing why pupillometry is not readily compatible with fast-paced VSL tasks and suggest alternative measures to use to examine learning during SL.

General Method

Participants

One hundred and sixty-eight adult subjects (54 in Experiment 2; 61 in Experiment 3; 53 in Experiment 4), all with normal or correct-to-normal acuity and color vision, participated for course credit or monetary compensation ($6). Five subjects in Experiment 2 were excluded due to corrupted file (1 subject) and poor eye tracking (4 subjects). Eight subjects in Experiment 3 were excluded due to corrupted files (4 subjects) and poor eye tracking (4 subjects), which was defined as missing more than 25% of their data. Thirteen subjects in Experiment 4 were excluded due to poor eye tracking (same definition as above). Thus, data analysis in Experiment 2 included 49 participants, in Experiment 3 included 53 participants and in Experiment 4 included 40 participants. Experiments were approved by the University's Institutional Review Board, and informed consent was obtained before beginning the study.

Stimuli

Stimuli consisted of 12 fractal images (from Karuza, Kahn, Thompson-Schill, & Bassett, 2017, **Figure 11**). To prevent luminance-related changes in pupil diameter, all visual stimuli were made isoluminant (equal luminance) to the background color of the experiment and to the experimental room. The original fractal images were in black and white. To make the images isoluminant, we first had to add color to it. We used Photoshop CS6 (Adobe Systems, San Jose, CA), imported the original images, created a new layer for each image and filled it with an

orange color (#F28C09), set the layer blending mode to difference and exported the new colorful images. Next, we used an isoluminance script provided by Daniel Swingley (personal communication, David Brainard and Daniel Swingley, 2016) which takes a standard RGB image, corrects the gamma, separates the image's chromaticity and luminosity, and sets the luminance to a constant value, thereby making the image isoluminant, without affecting the chromaticity. We selected a luminance value of 0.3 to match the lighting in the room and not distort the visual stimuli to the point of being unrecognizable.

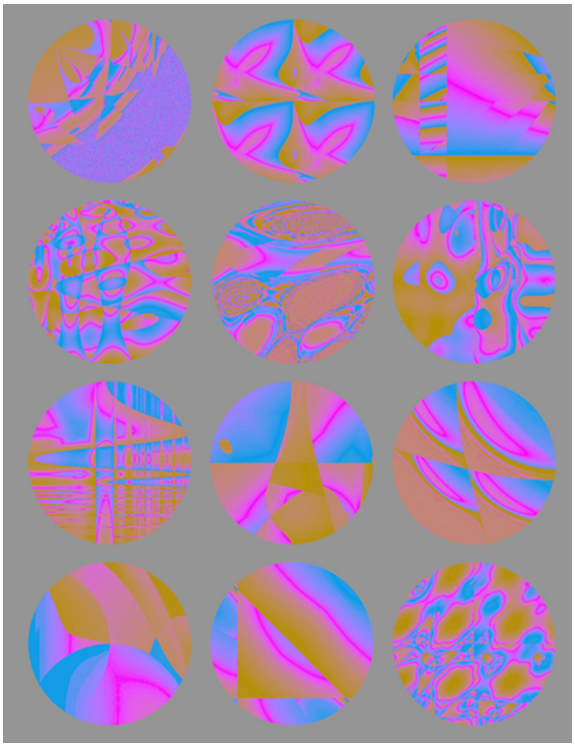


Figure 11. The 12 fractal images used in Experiment 2, Experiment 3 and Experiment 4.

Apparatus

Participants sat approximately 60 cm from the monitor and eye tracker (Eyelink 1000, SR Research, Ontario, Canada). We set the eye-to-track parameter to "right". The eye tracker

records pupil diameter in arbitrary units at 500Hz. Luminance in the room was measured using Light Meter, a mobile app used to measure light intensity (version 1.1, Elena Polyanskaya). The display monitor measured 34cm by 27cm and was facing the participant. The host monitor plus experimenting computer were in front of the experimenter. Before beginning the experiment, a five-point calibration was used. We performed calibration and validation for all participants.

Procedure

The basis for the current experiments is the standard VSL paradigm. In Experiment 2, participants viewed a stream of visual stimuli one at a time with no explicit instructions during the familiarization phase. Unbeknownst to the participants, the stream of visual stimuli was composed of four possible triplets. In other words, the three images that made up a triplet, always appeared one after another (e.g., ABC-DEF-GHI-ABC). During the test phase, participants were explicitly tested on how well they learned the four triplets using a two-alternative forced-choice (2AFC) test. In Experiment 3, participants viewed a stream of visual stimuli during familiarization composed of paired images and unpaired images (e.g., AB-X-CD-Y-EF-Z-AB-Y) and were given a simple recognition test during test phase. In Experiment 4, participants viewed a stream of visual stimuli composed of triplets identical to Experiment 2, followed by a random block composed of the same images but presented in a randomized order. There was no test phase in Experiment 4. For all three studies, we defined the period when there is a blank screen (i.e., before the image appears) as the anticipatory period and the period when the image is shown as the viewing period. The anticipatory period was 500ms and the viewing period was 750ms, for a total trial duration of 1250ms.

Preprocessing

Data analysis only included participants with good validation (i.e., the highest validation accuracy). Missing pupil size samples, identified as blinks, were linearly interpolated. To compare pupil diameter within and across participants, we calculated pupil change from baseline for both experiments. Using pupil change from baseline instead of raw pupil measurement accommodates variations in pupil diameter due to individual differences and is a standard approach for pupillometry data (Lavín et al., 2014; Mill, O'Connor, & Dobbins, 2016; Preuschoff et al., 2011). We defined baseline as the average pupil diameter during the anticipatory period of the first image in a group of three images. Next, the baseline was subtracted from each pupil diameter measurement in the group of three images and divided by the baseline to get the change from baseline. Then, we multiplied the number by one hundred to convert the decimal number to a percentage. We will use the term **pupil diameter** to refer to the original pupil diameter recorded in arbitrary unit, **pupil change** to refer to the change in pupil size from baseline. To characterize types of pupil change, **pupil constriction** refers to decrease in pupil change from baseline and **pupil dilation** refers to increase in pupil change from baseline.

Experiment 2: SL With Triplets And Two-Alternative Forced-Choice Test

In Experiment 2, we measured adults' pupil diameter as they participated in a typical VSL task. We modelled the paradigm after Turk-Browne et al. (2008), such that we included a passive exposure phase (i.e., familiarization phase) plus a surprise 2AFC test (i.e., test phase, **Figure 12**). This paradigm is suitable for analyzing individual differences because we can collect individuals' pupil diameter during familiarization and learning outcomes. To analyze pupil

diameter, we set the anticipatory period of the first image in a triplet or pair as baseline and calculated pupil change from baseline. We hypothesized that the pupil diameter will be different for each image in a triplet during the viewing period such that the pupil diameter will be the largest for the first image in the triplet (i.e., the least prediction image/largest prediction error) and smallest for the third image in the triplet (i.e., the most predictable) and the second image to fall in between. Furthermore, sensitivity to the order of the triplet images measured via pupil change will be related to performance during the test phase. We also calculated pupil change from baseline by using the anticipatory period of each image instead of the triplet. In Experiment 2 and subsequent experiments, we show that different selections of baseline can drastically alter results and lead to differences in pupil change unrelated to prediction.

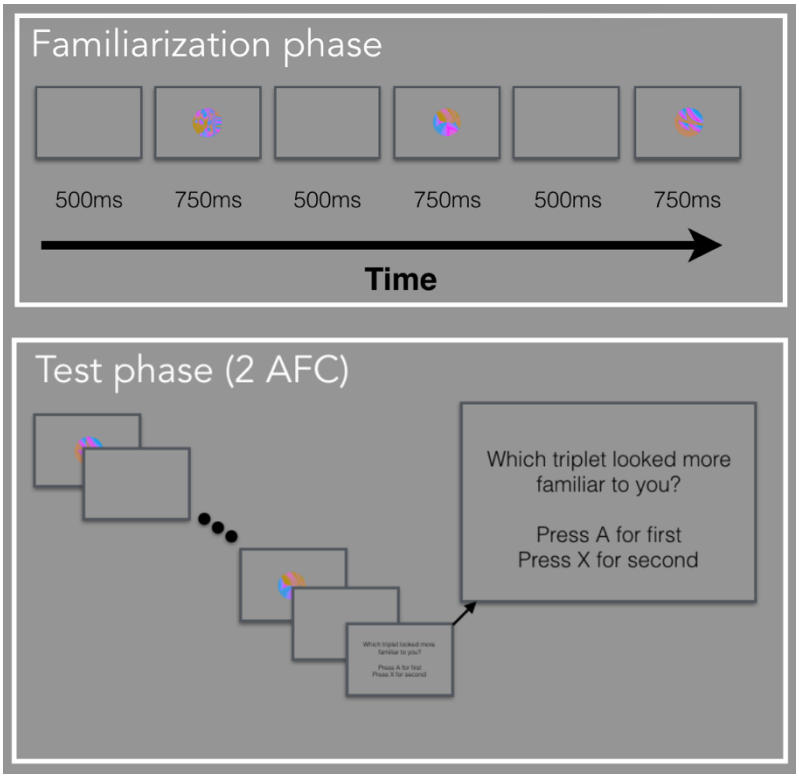


Figure 12. Schematic of VSL task in Experiment 2.

In addition, we complemented these analyses by computing Siegelman et al. (2018)'s online measure of SL:

average pupil change (1st position) – mean average pupil change (2nd and 3rd position)

Siegelman et al. (2018) conducted a self-paced VSL paradigm, where participants were exposed to a stream of images composed of triplets and asked to press a button to view the next image. Using this formula with log-transformed RTs, Siegelman et al. (2018) found that participants learned the triplet structure after only a small number of exposures. Furthermore, the RT difference for a predictable vs. unpredictable image was highly correlated with learning outcome measured during test. We computed Siegelman et al. (2018) online measure of SL using pupil change and analyzed its relationship to performance on the test phase. Lastly, it is important to mention that given this novel combination of using pupillometry during a VSL task, our overall statistical plan is to engage in a series of exploratory analyses in Experiment 2, uncorrected, and then attempted to replicate and extend them in independent datasets (Experiments 2 and 3).

Procedure

Familiarization phase. Participants watched a sequence of 288 images, presented in the center of the computer screen, one at a time. Unbeknownst to the participants, the sequence was composed of triplets, or sequences of three images that always appeared in the same order. We randomly assigned the 12 fractal images into 4 triplets (ABC, DEF, GHI, JKL) for each subject. The full sequence was generated by randomly interleaving 8 repetitions of each triplet with two

constraints (for a total of 24 repetitions of each triplet): (a) no triplet could be repeated sequentially (e.g., ABC-ABC), and (b) no pair of triplets could be immediately repeated (e.g., ABC-GHI-ABC-GHI).

Test phase. Participants completed 32 two-interval forced-choice test trials, judging the relative familiarity of (a) triplets versus (b) foil sequences of three familiar images that had never appeared sequentially during familiarization. The foil triplets were created by keeping each image's position the same, but changing the sequence within a triplet. For example, a foil could be made up of image sequence GBF, where each image within a triplet appears in the same position as they did before, but the sequence is new. The triplet and foil sequences were presented on each trial in the same manner as during familiarization, separated by a 1250ms pause. The order of presentation was randomly chosen with equal likelihood. Observers pressed "A" if the first sequence seemed more familiar and "X" if the second sequence seemed more familiar.

Results

Test Accuracy (Offline Measure of Learning)

As we are interested in individual differences, we first confirmed that the task elicited learning otherwise individual differences in the task would not be meaningful. The average accuracy score on the test phase was low (59%) but significantly above chance ($t(48) = 5.1, p < 0.001$). Above chance learning overall suggested that participants learned the triplet sequence (**Figure 13**).

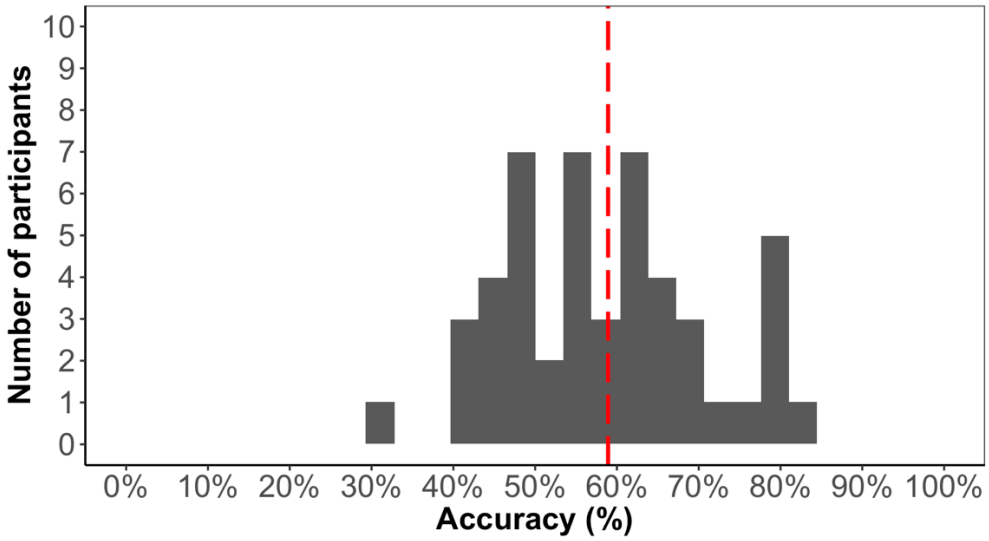


Figure 13. Distribution of test phase accuracy in Experiment 2.

Pupil Change Across Triplet Presentation (Online Measure of Learning)

Having confirmed overall learning in the task and a wide-range of accuracy scores (i.e., our offline measures) during test phase, we went on to analyze pupil diameter obtained during familiarization (i.e., our online measure). We sought to determine whether the magnitude of the pupil change varies according to participant accuracy, image position (i.e., if it's the first, second, or third image in a triplet), and number of times seen, and participant as a random effect. We ran two generalized linear mixed models (GLMMs), one using the average pupil change during the anticipatory period and the other using the average pupil change during the viewing period as the dependent variable for a single trial. We found that accuracy was not a significant predictor of pupil change for either anticipatory (β = -1.3e-2, p = 0.29) or viewing period (β = -2e-2, p = 0.7). To confirm whether the data favors the null hypothesis, we conducted a Bayes factor analysis comparing the GLMM with and without accuracy as a predictor. We found an extremely small Bayes factor of 0.02 for the anticipatory period and 0.05 for the viewing period, which suggests the data strongly favors the model without accuracy as a predictor. Overall, this

lack of relationship suggest that participants' pupil change during learning is not related to offline learning performance during the test phase.

Focusing on the image position within the triplet, we found that image position was a significant predictor of pupil change during the anticipatory period ($\beta = 6.3\text{e-}3, p < 0.001$) but not the viewing period ($\beta = 2.5\text{e-}3, p = 0.13$) (**Figure 14**). The more predictable an image, the smaller the pupil constriction during the anticipatory period. Specifically, the anticipatory period appeared to reflect anticipation of the upcoming image because the ordering of pupil constriction from smallest to largest was image one, two, three. Furthermore, we found a familiarity effect on pupil diameter such that the number of times you see the same stimulus significantly increases pupil change for both the anticipatory ($\beta = 4\text{e-}4, p < 0.05$) and viewing period ($\beta = 1.1\text{e-}3, p < 0.001$).

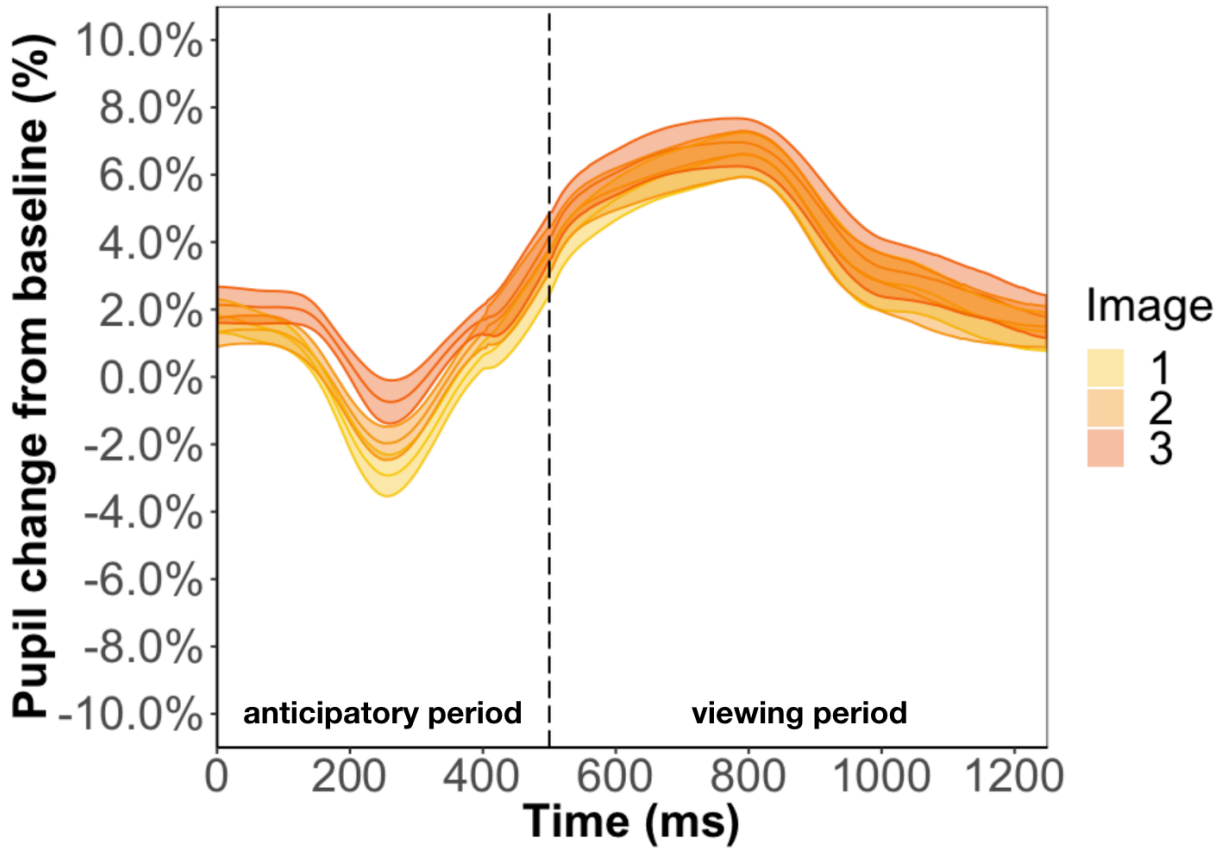


Figure 14. **Pupil timecourse during viewing of a triplet in Experiment 2**. The anticipatory period lasted for 500ms. The dashed vertical line indicates when the image appears on the screen. The viewing period lasted for 750ms, for a total trial duration of 1250ms. The images in a triplet are overlaid on top of each other to visualize the pupil change within a triplet.

Pupil Change Using Each Image As Their Own Baseline (Online Measure of Learning)

The analyses in the previous subsection found that pupil responses during the anticipatory period vary with triplet position. The data used for this finding employed a single baseline period per triplet (the anticipatory period for the first image). In this subsection, we reanalyze this data with each image's anticipatory period as their own baseline and find that this effect does not persist.

We ran two regression models, one using the average pupil change during the anticipatory period and the other using the average pupil change during the viewing period as the dependent variable. For the anticipatory period, no predictor was significant (image position: β = -1.7e-5, p = 0.27; accuracy: β = 3.1e-4, p = 0.63; familiarity: β = -1.2e-6, p = 0.52). For the viewing period, we found a familiarity effect (β = 7.3e-4, $p < 0.001$) but no effect of image position (β = 3.7e-4, p = 0.79) and accuracy (β = 9.6e-3, p = 0.85). These results demonstrate that a change in baseline selection can alter the results because the effect of position and familiarity during the anticipatory period is no longer significant. We will return to the question of proper baseline in later experiments.

Siegelman et al. (2018) Online Measure of SL

We found that the trajectory of the online measure for Experiment 2 was significant for the anticipatory period (r = -0.41, df = 22, p = 0.046) and approaching significance for the viewing period (r = -0.38, df = 22, p = 0.06). However, results revealed a non-significant relationship between Siegelman et al. (2018)'s online measure of SL and accuracy on test phase

(anticipatory period: $r = 0.16, df = 47, p = 0.26$, viewing period: $r = 0.025, df = 47, p = 0.87$). These results suggest that pupil change is not sensitive enough to reflect learning mechanisms in VSL.

Exploratory Analyses Suggested by Reviewers

We conducted exploratory analyses, as suggested by our Reviewers, to rule out alternative explanations. First, slower, tonic changes in pupil size were also investigated by investigating the unbaselined pupil size using the same GLMMs (Appendix B and Appendix C). We find the same pattern of effects as using the baselined pupil response which rules out the possibility that tonic pupil diameter is the better measurement of online learning. Second, a separate analysis using trials from the second half of the familiarization was conducted (Appendix D and Appendix E), to confirm that the mixed findings were not a result of including early trials, where learning might not have occurred. These results also converged the results of the main analyses suggesting that it is not slow or weak learning that is reducing the signal in the pupil response. Lastly, to confirm that the inclusion of poor learners did not obfuscate the results, we conducted identical analyses using the group of participants that performed above median, which had a median accuracy of 74%. This group also resulted in similar findings (Appendix F and Appendix G).

Discussion

In Experiment 2, we found that pupil change during VSL familiarization phase does not predict accuracy during VSL test phase. We also found that pupil change during the anticipatory period reflects image position in a triplet. Specifically, as participants view the least predictable image to the most predictable image, pupil constriction weakens. This change in pupil

constriction during the anticipatory period was an unexpected result for two reasons. First, the period which we find this shift in pupil change was unexpected because previous research found pupil change to reflect prediction error after stimulus presentation. In the current study, we found it before stimulus presentation and confirmed that it is not a result of participants not paying attention to the stimulus (Appendix H). Second, the direction of pupil change was opposite of what we hypothesized. Previous research has shown that pupil dilation reflects prediction error (Nassar et al., 2012; Zhang et al., 2018), such that the larger the prediction error the larger the pupil dilation. In this task, we found that pupil constriction might be related to prediction error.

We also found an influence of familiarity such that pupil change increased as the task progressed, which contradicts well-established findings that larger pupil dilation is a signal for surprise (Kloosterman et al., 2015; Nassar et al., 2012; Preuschoff et al., 2011). This unexpected finding might be caused by changes in tonic pupil diameter (i.e. baseline). Specifically, studies have shown a strong negative correlation between tonic pupil diameter and phasic pupil change (e.g. Gilzenrat, Nieuwenhuis, Jepma, & Cohen, 2010; Murphy, Robertson, Balsters, & O'connell, 2011), which is also confirmed in our data set ($r = -0.33$, $df = 13504$, $p < 0.001$). Therefore, because tonic pupil size is larger during the early trials of the experiment and decreases with time ($r = -0.63$, $df = 283$, $p < 0.001$), it likely explains why pupil change is small but increases as the task progressed. Overall, the results of Experiment 2 are mixed: A promising finding with respect to the pupil differences during the anticipatory period but it does not appear to be related to individual differences in performance at test. We move on to Experiment 3 to attempt to replicate the promising results to confirm whether they are veridical measures of VSL.

Experiment 3: SL With Paired Images And Recognition Test

In Experiment 2, the difference in pupil change occurred in an unexpected direction and location. We found pupil change increased with predictability and occurred during the anticipatory period before a stimulus (i.e., in the ISI following the previous stimulus). Moreover, pupil change during familiarization was not correlated with accuracy during test. For these reasons, we sought to replicate and extend these findings. We conducted another VSL study using paired and unpaired images, instead of triplets, and a yes/no recognition test during test phase instead of 2AFC discrimination test. Having both paired and unpaired images allowed us to look at how the pupil change for a predicting image, predicted image and a completely irrelevant image (i.e., unpaired). In particular, the unpaired image helps us disentangle pupil changes that are caused solely by familiarity versus changes in predictability arising from statistical learning. Furthermore, we used a familiarity test with the intention of making the test phase easier to increase the likelihood of learning and performance during the test phase. Again, Siegelman et al. (2018) online measure of SL was also calculated for both the anticipatory and viewing period.

Procedure

Familiarization phase. Participants watched a sequence of 288 images. The sequence was composed of pairs, or sequences of two images that always appeared in the same order, interleaved with unpaired (i.e., random) images. We randomly assigned the 12 fractal images into 4 pairs (AB, CD, EF, GH) and the remaining 4 images as the unpaired images (I,J,K,L), for every subject. The full sequence was generated by randomly interleaving 8 repetitions of each

paired and unpaired image with the constraint that no image could be repeated sequentially (e.g., AB-AB or AB-I-I).

Test phase. Participants completed 32 untimed test trials. Half of the test trials were foils and the other half were actual pairs. The foil pairs were created by keeping each image's position the same, but changing the sequence within a pair. For example, a foil could be made up of image sequence AF, where each image within a pair appears in the same position as they did before, but the sequence is new. Observers pressed "A" if the sequence was old and "X" if the sequence was new.

Results

Test Accuracy (Offline Measure of Learning)

The average accuracy score on the test phase was 53% (**Figure 15**), which is significantly above chance ($t(52) = 2.05, p < 0.05$), suggesting that participants learned the paired sequence during familiarization. However, the performance on this test is significantly lower compared to Experiment 2 ($t(99) = 2.5, p < 0.05$), suggesting that the change in familiarization from triplets to pairs and/or the change in test format made it more difficult to learn.

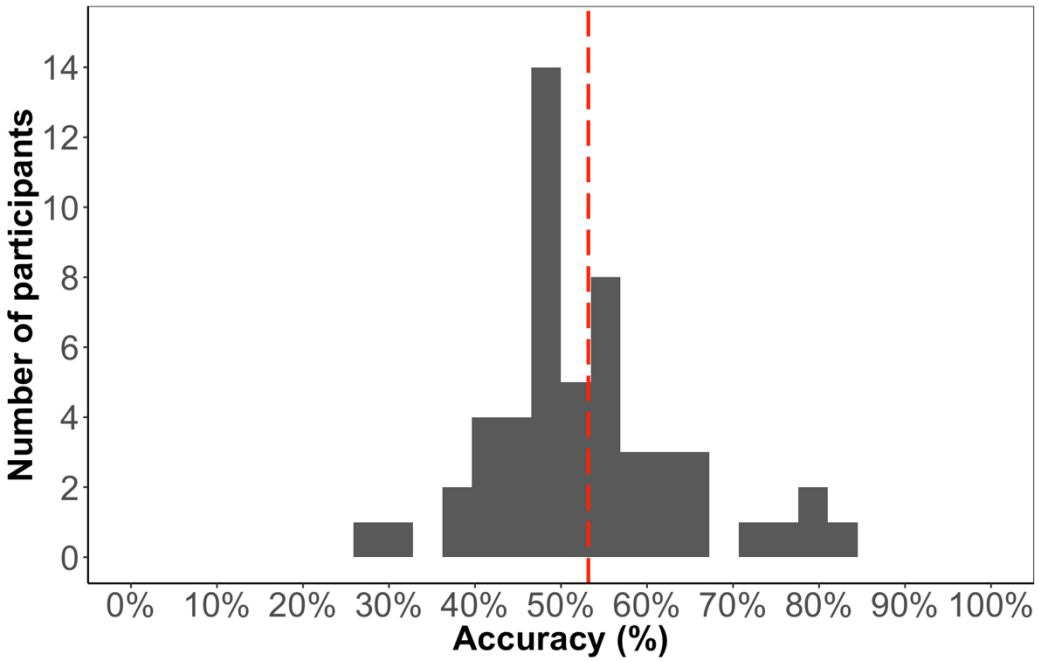


Figure 15. Distribution of test phase accuracy in Experiment 3.

Pupil Change Across Triplet Presentation (Online Measure of Learning)

The purpose of Experiment 3 was to replicate and extend the findings from Experiment 2. There are two principled ways to look at the data for this experiment with the goal of convergent findings. The two methods differ based on the trials included and how we baseline the pupil size. If pupil change is a valid online measure of SL, we should find the same effects across both methods.

For method one, we analyzed only the patterned sequence (i.e., paired images) and disregarded the unpaired images. Furthermore, we used the anticipatory period for the first image in a pair as baseline, which keeps the baseline selection consistent with Experiment 2. We ran two regression models: one using the average pupil change during the anticipatory period and the other using the average pupil change during the viewing period as the dependent variable. For both models, we included participant accuracy, image position (i.e., if it's the first or second

image in a pair), and number of times seen, as the independent variables and participant as a random effect.

Replicating the result of Experiment 2, we found that pupil change during the anticipatory (β = -9.7e-3, p = 0.35) and viewing period (β = -5.3e-3, p = 0.85) was not related to test phase accuracy. We also conducted a Bayes factor analysis comparing the GLMM with and without accuracy as a predictor. We found an extremely small Bayes factor of 0.01 for the anticipatory period and 0.03 for the viewing period, which suggests the data strongly favors the model without accuracy as a predictor. Inconsistent with the results of Experiment 2, we found image position to be a significant predictor of pupil change during the anticipatory period (β = 4.9e-3, p < 0.01) and viewing period (β = 7.4e-3, p < 0.01) (**Figure 16**). In other words, the predicted image had a larger pupil change than the predicting image for the entire trial instead of just the anticipatory period, which is what we found in Experiment 2. Furthermore, we found a familiarity effect on pupil change during the viewing period (β = 3.4e-4, p < 0.05), but not anticipatory period (β = 2.4e-5, p = 0.80), whereas in Experiment 2 we found a familiarity effect for both periods.

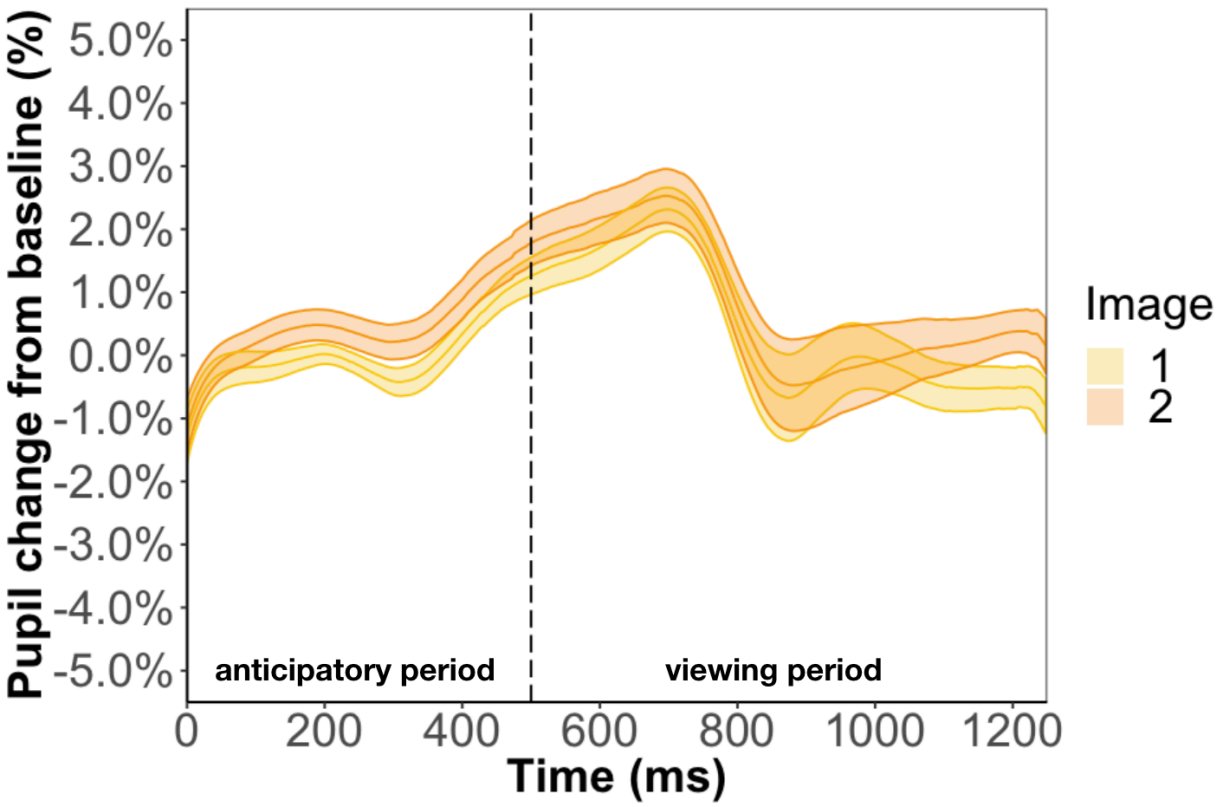


Figure 16. **Pupil timecourse in Experiment 3 using the anticipatory period of the first image in a pair**. The unpaired image was removed from analysis.

Method two was a more rigorous analysis and involved keeping the unpaired image so that we could disentangle the effect of familiarity from prediction. For this method, the baseline was the anticipatory period of the most recent unpaired image. However, this baseline selection led to undistinguishable pupil changes between the three types of images (**Figure 17**). Specifically, the anticipatory pupil difference we found in Experiment 2 goes away suggesting that this effect is not a result of predictability per se but baseline selection. We tackle this issue more directly in Experiment 4.

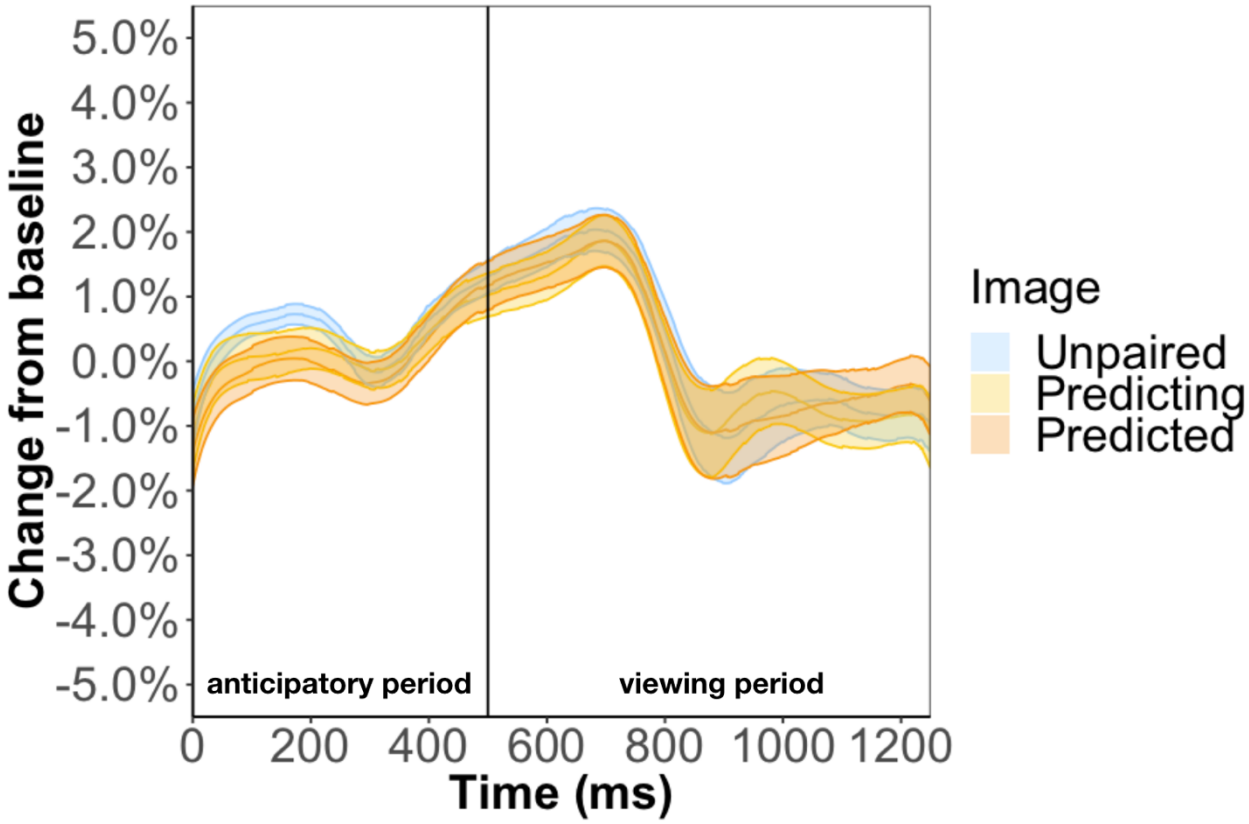


Figure 17. Pupil timecourse in Experiment 3 using the anticipatory period of the most recent unpaired image as baseline.

Pupil Change Using Each Image As Their Own Baseline (Online Measure of Learning)

We consider method one results using a different baseline selection where each image's anticipatory period is their own baseline. We did not conduct this analysis for method two because of the lack of significant results. For the anticipatory period, no predictor was significant (image position: $\beta = 5e\text{-}6, p = 0.52$; accuracy: $\beta = 1e\text{-}5, p = 0.88$; familiarity: $\beta = 1.6e\text{-}7, p = 0.66$). For the viewing period, we found a familiarity effect ($\beta = 3.6e\text{-}4, p < 0.01$) and image position ($\beta = 7.1e\text{-}3, p < 0.01$) but no effect of accuracy ($\beta = 2.2e\text{-}3, p = 0.93$). Again, altering the baseline period drastically altered the results.

Siegelman et al. (2018) Online Measure of SL

We calculated Siegelman et al. (2018) online measure of SL using just the paired images. We found that the trajectory of the online measure was not significant for both the anticipatory ($r = 0.006$, $df = 34$, $p = 0.97$) and viewing period ($r = -0.16$, $df = 34$, $p = 0.35$). Furthermore, results revealed a non-significant relationship between online measure of SL and accuracy on test phase (anticipatory period: $r = 0.14$, $df = 50$, $p = 0.31$, viewing period: $r = 0.03$, $df = 50$, $p = 0.84$). These results provide further evidence that pupil change is not reflecting the learning mechanisms supporting VSL.

Discussion

In Experiment 2, the difference in pupil change occurred in an unexpected direction and location. We found pupil change increased with predictability and occurred during the anticipatory period before a stimulus (i.e., in the ISI following the previous stimulus). Moreover, pupil change during familiarization was not correlated with accuracy during test. For these reasons, we sought to replicate and extend these findings.

In Experiment 3 we conducted another VSL study using paired and unpaired images, instead of triplets, and a yes/no recognition test during test phase instead of 2AFC discrimination test. The changes in experimental procedure led to one consistent result and several inconsistent results. Using the first image of the paired images as baseline and removing the unpaired image, we replicated the main finding from Experiment 2: We found that pupil change during the anticipatory period reflected image position in a pair with a greater pupil change during more predictable pictures. Furthermore, the differentiable pupil changes across position, did not predict performance during test. Focusing on the inconsistent results, first, while we find an

effect of position in the anticipatory period, we also found this effect during the viewing period which was not present in Experiment 2. Second, we found the familiarity effect during the viewing period but not anticipatory period. The last and most concerning result was when we used the anticipatory period of the most recent unpaired image as baseline, the anticipatory pupil difference we found in Experiment 2 disappeared. This inconsistent pattern of results led us to be skeptical that the pupil differences we found during the anticipatory period in Experiment 2 and Experiment 3 is a valid online measure of SL. Therefore, we conducted Experiment 4 to investigate whether pupil changes in the anticipatory period is a result of prediction or baselining.

Experiment 4: VSL Using Triplet And Random Blocks

In Experiment 4, we investigated whether pupil changes in the anticipatory period are reflective of online learning processes or are an artifact of our baselining procedure. In other words, is the pupil change during anticipatory period related to participants' learning (e.g., prediction for an upcoming image)? Or is it a result of simply choosing baseline to be the anticipatory period of the first image in a triplet or pair? To distinguish between the two alternatives, we used a within subject design to measure pupil change during a triplet block and a random block. Our hypothesis is if the differences in pupil change is due to stimulus prediction, then we should only see pupil difference in the anticipatory period during the triplet block. However, if the differences in pupil change is due to baselining then we should see the difference in both triplet and random block. Siegelman et al. (2018) online measure of SL was also calculated for both blocks.

Methods

Procedure

The experiment was made up of two blocks: triplet and random. All participants first completed the triplet block, followed by the random block. We did not counterbalance the blocks, because successively presented sets of stimuli can bias or reduce learning (Gebhart, Aslin, & Newport, 2009). Participants watched a sequence of 480 fractal images (240 images per block). In the triplet block, the sequence followed the same rules as the familiarization phase of Experiment 2. In the random block, the full image sequence was presented in a complete randomized order, with the exception that no image could be repeated sequentially. We did not include a test phase for Experiment 4.

Results

Pupil Change Across Triplet Presentation (Online Measure of Learning)

We ran two regression models to look at what predicts pupil change during the anticipatory period. One model used average pupil change during the anticipatory period of the triplet block and the other model used the average pupil change during the anticipatory period of the random block (**Figure 18**). If the pupil change in anticipatory period in Experiment 2 is due to stimulus prediction, we expect to find image position as a significant predictor for the triplet block but not the random block. If the pupil change in anticipatory period is due to temporal prediction, we expect to find image position as a significant predictor for both the triplet and random block. To test what was the cause of the pupil change, we regressed average pupil change on image position, and number of times seen for both models. In terms of image position, because there is no picture predictability in the random block we simply grouped three

consecutive images together to form a “random triplet”, whereas the group of three images in the triplet block was a “patterned triplet”.

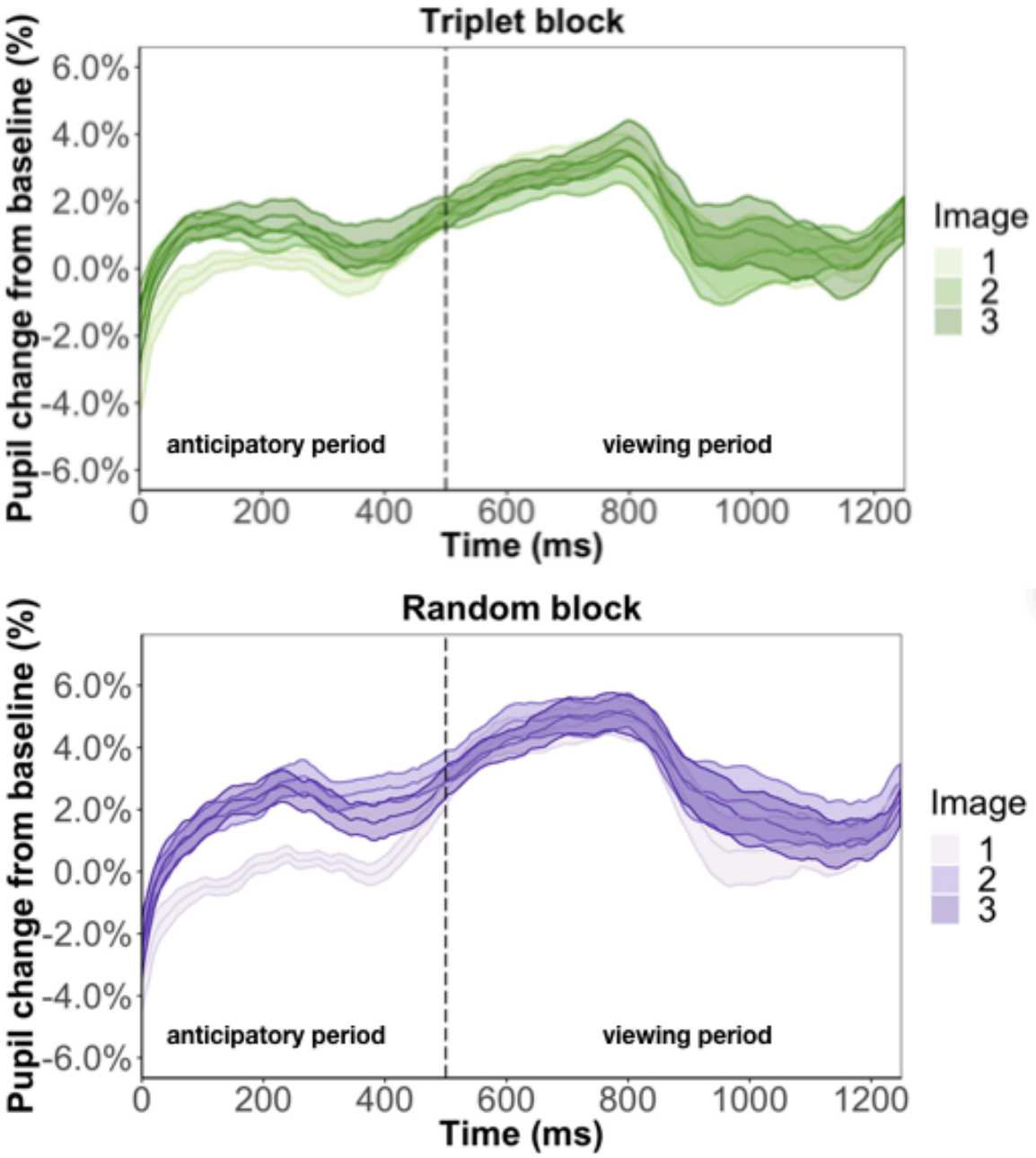


Figure 18. **Pupil time-course during viewing of a triplet in Experiment 4**. The anticipatory period lasted for 500ms. The dashed vertical line indicates when the image appears on the screen. The viewing period lasted for 750ms, for a total trial duration of 1250ms. The images in a triplet are overlaid on top of each other to visualize the pupil change within a sequence of three images.

We found that image position was a significant predictor of pupil change during the anticipatory period for both the triplet block ($\beta = 5.3e\text{-}3, p < 0.001$) and random block ($\beta = 8.6e\text{-}3, p < 0.001$). In other words, pupil dilation was smallest for the first image in a triplet and largest for the third image in a triplet, with the second image falling in the middle, for both the triplet block and random block. This suggests that the pupil difference during the anticipatory period is likely an effect of where we chose to baseline, instead of stimulus prediction, because we find this difference in the random block, where the stimuli are not predictable. Furthermore, we found a familiarity effect on pupil change during the anticipatory period for the triplet block ($\beta = 6e\text{-}4, p < 0.01$) but not random block ($\beta = 7.2e\text{-}5, p = 0.79$). This suggests that the unstructured nature of the random block caused participants to not recognize the individual stimuli and view each appearance as it was their first time seeing the image.

We also ran two regression models to look at what predicts pupil change during the viewing period. Replicating Experiment 2, we found evidence for a familiarity effect on pupil change during the triplet block ($\beta = 7.5e\text{-}4, p < 0.01$) and we did not find effect of image position during the viewing period ($\beta = 6.1e\text{-}4, p = 0.71$). This result suggests that the inconsistent findings from Experiment 3 is caused by changing the triplets to pairs.

Pupil Change Using Each Image As Their Own Baseline (Online Measure of Learning)

Using each image's own anticipatory period as baseline, we found that during the anticipatory period, image position was almost a significant predictor ($\beta = -1.86e\text{-}5, p = 0.088$) but no familiarity effect ($\beta = -1.5e\text{-}6, p = 0.33$). However, for the random block, image position was also almost a significant predictor ($\beta = -2.5e\text{-}5, p = 0.07$) and there was a familiarity effect ($\beta = -4.6e\text{-}6, p = 0.02$). During the viewing period, there was no significant predictors for either blocks.

Siegelman et al. (2018) Online Measure of SL

We found that the trajectory of the online measure was not significant for the triplet block (anticipatory period: $r = -0.40$, $df = 18$, $p = 0.08$, viewing period: $r = -0.23$, $df = 17$, $p = 0.34$) or random block (anticipatory period: $r = 0.03$, $df = 18$, $p = 0.9$, viewing period: $r = 0.14$, $df = 18$, $p = 0.55$). Importantly, when comparing the learning trajectories between the triplet and random block, we found no significant interaction. This lack of difference suggests that participants did not differ in how well they learned the statistical structure for an input with structure versus an input without structure.

Discussion

In Experiment 4, we found pupil change during the anticipatory period reflects image position for both the triplet and random block. This result suggests that the pupil change during the anticipatory period in both Experiment 2 and Experiment 3 is a result of baselining rather than stimulus predictability. Based on the results of Experiment 4, we can reinterpret the pupil difference between image position during the anticipatory period of Experiment 2 and 2 as purely due to baseline selection, not predictability.

General Discussion

In three experiments, we measured participants' pupil diameter using VSL. In Experiment 2, we employed a VSL task using images grouped into triplets during familiarization and 2AFC task during test. In Experiment 3, we employed a VSL task using paired and unpaired images during familiarization and a familiarity task during test phase. In Experiment 4, we employed a VSL task without the test phase and showed participants a sequence of images first

composed of triplets (i.e., triplet block) followed by a sequence of random images (i.e., random block). Overall, we found that pupil diameter is not a good online measure to use during SL given that it does not predict offline measures of learning (i.e., performance during test phase) and varies substantially according to baseline selection (for summary of results see **Table 1**). Finally, drawing from previous pupillometry studies looking at prediction and prediction error, we propose that these negative effects arise from the fact that the pupil response unfolds too slowly to be compatible with the fast nature of VSL.

A.

	Accuracy		Image position		Familiarity	
	anticipatory	viewing	anticipatory	viewing	anticipatory	viewing
Experiment 2 (Pupil Change Using Triplet)						
Experiment 3 (Pupil Change Using Triplet)						
Experiment 4 (Triplet Block: Pupil Change Using Triplet)	n.a	n.a				
Experiment 4 (Random Block: Pupil Change Using Triplet)	n.a	n.a				

B.

	Accuracy		Image position		Familiarity	
	anticipatory	viewing	anticipatory	viewing	anticipatory	viewing
Experiment 2 (Pupil Change Using Image)						
Experiment 3 (Pupil Change Using Image)						
Experiment 4 (Triplet Block: Pupil Change Using Image)	n.a	n.a				
Experiment 4 (Random Block: Pupil Change Using Image)	n.a	n.a				

Table 1. **Overview of the two different baseline selection results for all experiments.** Red indicates the result was not significant, green indicates the result was significant, and blue suggests approaching significance. A. Pupil change was calculated for each triplet by selecting a single baseline period (the first image's anticipatory period) per triplet. B. Pupil change was calculated by using each image's anticipatory period as their own baseline.

First, if pupillometry was a good measure of prediction and learning during VSL we would expect it to be sensitive enough to distinguish the participants that learned the patterned sequence from those that didn't. However, in Experiment 2 and Experiment 3, we found that pupil change during familiarization (i.e., learning) did not predict performance during test phase. One possibility for this is that the pupil response does not reflect the relevant online learning mechanisms. We argue that this is not likely to be the case: As reviewed above, pupillometry has been shown to be sensitive to processes that are part of prediction such as prediction error (Koenig et al., 2018; Nassar et al., 2012; Zhang et al., 2018) and updating (O'Reilly et al., 2013). Thus, it is likely that the pupil response is reflecting relevant online processes during learning.

Second, pupil diameter is not suitable for SL because where we chose to set baseline drastically altered pupil change. In Experiment 2, we found that pupil change during the anticipatory period reflected image position in a triplet such that the pupil constriction was smallest for the third image (i.e., the most predictable image) in a triplet and largest for the first image (i.e., the least predictable image) in a triplet. In Experiment 3, we used paired and unpaired images instead of triplet images because we wanted to disentangle pupil change associated with familiarity from prediction. However, data analysis using the anticipatory period of the unpaired image as baseline led to undistinguishable pupil responses. On the other hand, when we removed the unpaired image from data analysis and used the anticipatory period of the first image in a pair as baseline, we found the same pupil difference in the anticipatory period as Experiment 2. To determine whether this pupil difference during the anticipatory period was a result of stimulus prediction or baseline selection we conducted Experiment 4. The results of Experiment 4 revealed a significant difference in pupil change during the anticipatory period in both the triplet block and random block (i.e., no predictive processing) suggesting the underlying cause for this shift baseline selection rather than stimulus prediction.

One might suggest that there is no need to baseline pupil diameter; however, there are two reasons why it is very important to baseline pupil size. The Eyelink 1000 eyetracker records the pupil diameter in arbitrary units in the range of 400 to 16,000 units to represent pupil diameter with large differences across individuals and within individuals throughout the experiment. Without baselines, this substantial variation will likely wash out any systematic differences in pupil size induced by your experimental task. Moreover, this arbitrary unit can be tricky to convert to mm because the coefficient of proportionality depends on the experimental layout (i.e., the relative positions of the camera, eye and monitor) (Hayes & Petrov, 2015). Using

an artificial eye, and selecting “Left” on the Eyelink 1000 eyetracker, Hayes and Petrov (2015) found that pupil diameter recordings tend to increase as gaze position moves from left to right throughout the middle of the screen. Using data we collected for a separate infant study in which we specifically counterbalanced which eye we collected data from, we found the same pattern of pupil deviation for the left eye, but for pupil diameter that was recorded using the “Right” eye we found the opposite trend (personal communication Gabriel Xiao, **Figure 19**). In other words, when using a remote eye-tracking system, where the participant is sitting still in front of the computer screen and the eye-tracker is attached below the computer screen, pupil size becomes distorted as participants’ eyes scan the screen. Furthermore, this pupil distortion is not the same in both eyes and therefore, it is imperative to calculate percent change from baseline.

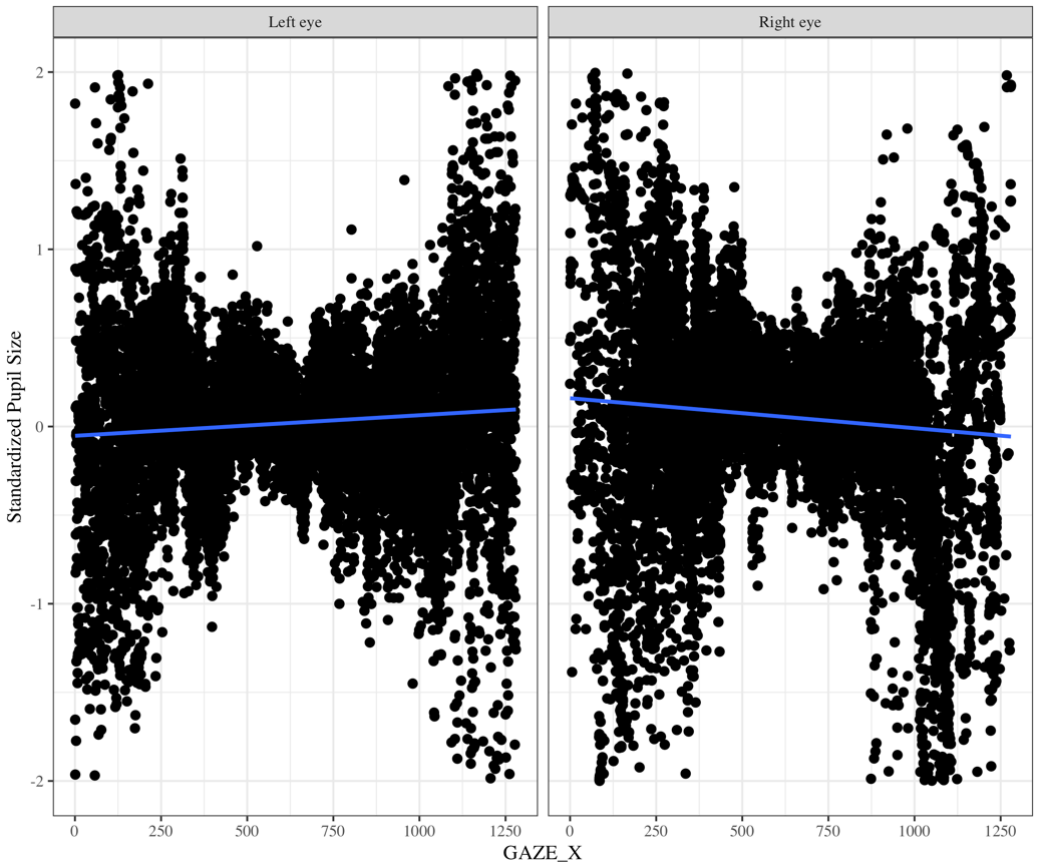


Figure 19. Changes in pupil diameter based on X coordinate eye position for left eye (left panel) and right eye (right panel) when using a remote eye-tracking system. Pupil diameter was z-scored to the infant's average pupil diameter for the entire experiment.

In addition, depending on the type of analysis, baselining is important because the pupil diameter tends to increase throughout the experiment due to drift. It is important to note that even though we baselined our pupil diameter for all our studies, we still found increased pupil change throughout the experiment. Results from our studies demonstrate the influence of familiarity on pupil change. For all three experiments, we found results supporting research suggesting that demonstrates larger pupil dilation for more familiar stimuli (Kafkas & Montaldi, 2012, 2015). But this increase in pupil change could have been even larger if we didn't baseline. Therefore, baselining can help reduce the effect of drift on pupil diameter.

However, deciding which portion of the task to use for baseline is difficult. Generally, which portion of the task you choose to be your baseline depends on the task. Some studies have used the beginning of the experiment (Tsukahara, Harrison, & Engle, 2016), some studies use beginning of every trial or before stimulus onset (Byers-heinlein, Morin-lessard, & Lew-williams, 2017; Koenig et al., 2018; Murphy, Van Moort, & Nieuwenhuis, 2016; Wetzel, Buttelmann, Schieler, & Widmann, 2016) some studies use the average pupil diameter for the entire experiment as baseline (Geng, Blumenfeld, Tyson, & Minzenberg, 2015) and some compute z-score pupil size for the entire experiment (Kang, Huffer, & Wheatley, 2014; Kang & Wheatley, 2015). Ideally, one should baseline for each stimulus to avoid creating artifacts based on base-lining but this isn't likely to happen for VSL because of the type of effects researchers are interested in. In other words, baselining pupil diameter for each trial can remove experimental effects.

Beyond the issues of baselining, we propose that there is a more fundamental issue: Pupillometry may not be compatible with VSL due to the slow nature of the pupil change and the fast nature of SL tasks. Many studies that have revealed that the pupil response reflects processes like prediction error find these effects persisting many seconds after stimulus presentation. Indeed, studies that have included a long enough ISI for the pupil change to return to baseline (Lavín et al., 2014; Zhang et al., 2018) have been successful in using pupillometry in their respective task. However, all VSL tasks, however, present each stimulus for only 1000ms or less, with an even shorter ISI, resulting in an extremely short trial (e.g., Abla & Okanoya, 2009; Kirkham, Slemmer, & Johnson, 2002; Turk -Browne et al., 2005) that prevents pupil change from fully returning to baseline. Longer ISIs in VSL tasks may substantially change the mechanisms supporting VSL and/or result in no learning. Therefore, we propose that the timing

required by VSL studies (and SL studies more broadly) is not readily able to capture the aspects of the pupil response that reflect the relevant online processes.

This timing issue is reminiscent of the timing constraints of fMRI. FMRI has been used to investigate the mechanisms supporting VSL (e.g., Turk-Browne, Scholl, Chun, & Johnson, 2010) and has incorporated the longer timing necessary (e.g., 3-6 second ISIs). Using this design of fMRI style design may be a future avenue for merging pupillometry and VSL. However, given the substantial changes in these paradigms compared to standard VSL tasks and the difficulty in investigating VSL in developmental and differently-abled populations with these paradigms (i.e., because of the slow rate of stimuli and need for longer exposures), it raises the question of how useful pupillometry is as an online measure of VSL.

While pupillometry has been a highly fruitful online measure during a variety of tasks, we propose that it is severely limited in its application to investigating SL. Moving forward, researchers interested in the online mechanisms of SL can continue to experiment with the available methods. For adults, certainly fMRI is a fruitful alternative and the need for slower timing is complemented by the acquisition of a signal reflecting a myriad of systems in the brain. A review by Karuza, Emberson and Aslin (2014) argues that fMRI is an excellent online measure of learning as it can measure fluctuations in neural activity as learning unfolds over time, thereby allowing researchers to tap into the process through which internal representations undergo changes. Behaviorally, we have two RT measures available: click detection and self-paced paradigm. Comparing the two available alternatives, the self-paced paradigm seems to be the better option because there are less possible confounds given there isn't a cover task. Overall, current neuroimaging and behavioral methods still have the ability to expand our current

knowledge of SL and reveal the online mechanisms that occur during learning in SL and we propose they are superior to the use of pupillometry for VSL.

Chapter IV: Exploring prediction and prediction error using eye movements, pupillometry and learning models

Developmental research has revealed a lot about what infants can learn but much less about how they learn. For example, we know a lot about the types of input (e.g. *visual*: Bulf, Johnson, & Valenza, 2011; Tummeltshammer, Amso, French & Kirkham, 2016; *auditory*: Saffran, Aslin & Newport, 1996; Yu & Smith, 2007) that infants can learn from, but we don't know how infants are processing that input during learning. It is important to understand infant learning mechanisms because for many aspects of child development, such as language, the timing of experience, and relatedly how an infant learns from that experience, is crucial (Smith & Yu, 2013; Wells, Christiansen, Race, Acheson & MacDonald, 2009). Moreover, some children might receive the same (statistical) input but still struggle to learn and develop based on that experience (Elsabbagh et al., 2013; Evans, Saffran, & Robe-Torres, 2009; Mayo & Eigsti, 2012). For these reasons, it is important to understand the mechanisms of infant learning so that we may assess the efficacy of these mechanisms and intervene when these mechanisms show signs of irregularities. The current study focuses on one possible mechanism involved in learning early in life: prediction and, relatedly, prediction error.

Prediction is the ability to use one's experience and anticipate future, proximal input. Research with infants has established that prediction, as measured in numerous ways, starts very early in life (Aslin, 1981; Canfield & Haith, 1991; Háden, Németh, Török, & Winkler, 2015) and in various domains (*action*: Falck-Ytter, Gredebäck & von Hofsten, 2006; Gredebäck, Lindskog, Juvrud, Green & Marciszko, 2018; *language*: Kutas, Delong & Smith, 2011; Ylinen, Bosseler, Junttila & Huotilainen, 2016; *prosocial behavior*: Jin, Houston, Baillargeon, Groh & Roisman, 2018; *visual*: Romberg & Saffran, 2013; Rose, Feldman, Jankowski & Caro, 2002). Recent work

has extended these findings to show that prediction can modulate perceptual cortices in infancy suggesting that it can directly impact the processing of sensory input (Emberson, Richards, & Aslin, 2015; Kouider et al., 2015) rather than through facilitating eye movements. Furthermore, predictions can be correct or incorrect, and each type of prediction differentially influences learning. Research with adults has established a mechanistic role of prediction error (i.e. incorrect predictions) for learning (Bubic, Von Cramon & Schubotz, 2010; Den Ouden, Friston, Daw, McIntosh & Stephan, 2009; Den Ouden, Kok & De Lange, 2012). However, despite the decades of research investigating infants' ability to predict, there is little work directly relating prediction error to learning in infancy.

In a seminal study, Stahl and Feigenson (2015) found that after infants experienced a situation that is vastly different from their everyday lives (e.g., a car hovering in mid-air), they learn better about that object (e.g., the car). The authors invoke prediction error to explain this phenomenon because witnessing a hovering car violates an infants' knowledge of gravity. While this is plausible, it is unclear that prediction error, as it is established in models of error-driven learning, is what's at play here. Moreover, other studies have found that toddlers learn better when viewing a predictable event and not unpredictable event (Benitez & Saffran, 2018). Thus, there are contrary findings about the relationship between prediction, prediction error and learning in early development and previous studies haven't quantified prediction and prediction error systematically.

Importantly, there are well-validated and explicitly formulated models of learning that include prediction and prediction error, that may be used to try to understand this relationship in early development. The Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972) is a well-established associative learning theory that explicitly includes prediction and prediction error.

The RW model states that prediction errors will result in corresponding changes in associations and no change will occur if the error is zero. Another well-established learning model, the Pearce and Hall model (1980), states that prediction error drives attentional learning. For example, Hogarth, Dickinson, Austin, Brown, & Duka (2008) found that participants had longer fixation time to a cue that was associated with an uncertain auditory stimulus (i.e. high prediction error) compared to cues that were always associated with the presence or absence of an auditory stimulus (i.e. no prediction error). The common thread is prediction errors create the opportunity to update one's predictions or associations based on new information.

Given the behavioral evidence of a possible error driven learning mechanism present in infants (Stahl & Feigenson, 2015; and see related work linking prediction and learning: Benitez & Saffran, 2018; Benitez & Smith, 2012), we designed a study that incorporates both behavioral measures and computational learning models to systematically study the relationship between prediction error and learning (formation of prediction). The combination of behavioral data and computational modelling is uncommon in developmental psychology, but we believe this approach will lead to a more systematic and mechanistic understanding of how infants learn. To this end, the goal of the current study is to use the Visual Expectation Paradigm (VExP) and study how prediction error (i.e. incorrect eye movements) influences learning, by incorporating pupillometry, the measurement of pupil size.

In 1988, the Visual Expectation Paradigm (VExP) was introduced by Haith, Gazan and Goodman (1988) to study the emergence of prediction in very young infants using eye movements. In a standard VExP, infants watched individual visual stimuli (i.e. targets) appear in a simple left-right alternating sequence. The targets on one side followed a predictable order (targets appeared following a pattern) and the targets on the other side followed an unpredictable

order (targets appeared randomly). Each trial began with a 1000 ms delay followed by the target for 700 ms. The results showed that infants were more able to initiate eye movements during the predictable sequence, such that they will move their eyes during the delay to the side where the next target will appear, than the unpredictable sequence. We will refer to these as **anticipatory eye movements** (AEMs). In addition, infants also produced faster **reactive eye movements (REM)** to the target location, which are eye movements that occur after the target appears, during the predictable sequence than the unpredictable sequence. Both of these findings helped to establish that infants were able to predict upcoming stimuli and this prediction was reflected in their eye movement to the stimuli. However, no studies to this date have examined how error or incorrect AEMs influences performance in VExP. Thus, in order to systematically study prediction error, we modified the VExP and replaced the unpredictable sequence with a semi-predictable sequence designed to elicit prediction error.

To measure infants' experience of prediction error, we recorded infants' pupil size. Recent studies have established that pupil size closely corresponds to prediction error in infancy. Gredebäck et al. (2018) showed 6-months-old infants' videos of a person eating and found that some 6-months-old predicted another person's actions by looking at a person's mouth before the spoon reaches their mouth. Interestingly, the infants that made these correct anticipations had larger pupil change when they later saw a violation of feeding actions or experienced a prediction error. Moreover, infant pupil change is not simply an all-or-nothing error response but also scales with prediction error. For example, Tamási, McKean, Gafos, Fritzsche & Höhle, (2016) showed 30-months-old children an image and an auditory label that matched the image. The labels were words that 30-months-old children knew. The researchers manipulated the auditory label such that there were different levels of label violation (i.e., different amounts of

prediction error). The findings show that there was an overall pupil difference between correct and incorrect labels, with larger pupil change for incorrect labels. Furthermore, the incorrect labels that deviated more from the correct label had larger pupil change than incorrect labels that deviated less.

One can also combine pupil size and computational learning models to determine how closely the pupil size reflects prediction error. Zhang, Jaffe-Dax, Wilson & Emberson (2018) used pupillometry data in conjunction with error-driven learning models (RW model) in both adults and infants. All participants completed the identical implicit learning task designed to elicit associations between sounds and images. After a period of familiarization, participants also viewed error trials (when the sound was not followed by a corresponding image). By analyzing the data using computational models, the trial-by-trial changes in a participant's prediction or association and prediction error were revealed. They found that infant and adult's pupil response to stimuli significantly correlated with the model's estimation of prediction error and were able to show convergence of the participants' prediction and prediction error to the structure of the task.

We modelled the data in the present study using two well-established error driven learning models, the Pearce-Hall (PH) model and Rescorla-Wagner (RW) model, to quantify the relationship between prediction error and learning. The VExP can be used in conjunction with the PH model and RW model because infants can form an association between the spatial location (i.e. cue) and the target (i.e. outcome) and ideally make dissociable predictions about how the nature of the sequence (Predictable vs. Semi-Predictable) affects oriented to a particular spatial location. Previous work with adults has dissociated these two learning models using pupil responses in an associative learning task. Koenig, Uengoer and Lachnit (2018), conducted

three experiments with different learning conditions in adults and examined the relationship between pupil response, prediction error and these two learning models. Overall, Koenig et al. (2018) found evidence that pupil responses fit best with the predictions of the Pearce-Hall model. This finding is similar to other studies that have established an attentional bias for uncertain cues during learning using dwell time (Beesley, Nguyen, Pearson, & Le Pelley, 2015; Hogarth et al., 2008) and emphasizes the importance of attention for learning.

With respect to the current study, the PH model predicts that the amount of attention given to a cue will be determined by the prediction error. Specifically, the model predicts that the spatial location that's more uncertain (i.e. Semi-Predictable Sequence) will attract more attention than the spatial location that is consistently reinforced (i.e. Predictable Sequence). Therefore, if infants are exhibiting a preference, measured as more correct AEMs and faster REMs for the semi-predictable sequence, the PH model should be the best fit for the pupil data. Interestingly, attentional learning is also supported by behavioral research with infants demonstrating that 7 and 8 months old pay more attention to events that have an intermediate level of complexity compared to events that are overly simple (i.e. predictable) and overly complex (i.e. unpredictable, Kidd, Piantadosi & Aslin, 2012; 2014).

By contrast, the RW model states that the sign of prediction errors will determine the kind of learning that takes place. If the error is positive, or the outcome is better than predicted, the association between the cues and the outcome will strengthen. If the error is negative, or the outcome is worse than expected, the association between the cues and the outcome will weaken. With respect to the current study, the predictable sequence will have more positive associations between cue and outcome compared to the semi-predictable sequence because the association between the cue and outcome is consistently reinforced. Therefore, if infants are exhibiting more

signs of learning (i.e. more correct AEMs and faster REMs) for the semi-predictable sequence, the RW model should be the best fit for the pupil data.

The comparison of the two learning models can reveal which error driven theory best accounts for infant learning. If the PH model is a better fit, this suggests that prediction error is analogous to attention and if the RW model is a better fit, this suggests that prediction error plays a much more direct role in learning. As mentioned earlier, Zhang et al. (2018) found that pupil responses in infants and adults were significantly predicted by the RW model, though only one model was included, and the paradigm was much simpler. Overall, the results of this study will allow us to answer whether infants' experience of prediction error allows for a type of error-driven learning that is qualitatively (i.e. PDR) and quantitatively (i.e. computational modelling) similar to adults.

It is important to note that there are a number of key differences between Koenig et al. (2018) and the current study: Most notably, we are investigating these relationships in human infants and incorporating measures of prediction and learning in addition to prediction error. Second, Koenig et al. (2018) measured pupil change after the cue but before the outcome appeared and argued that this period, normally seen as a period of anticipation, changed as a function of error and is therefore more reflective of error than anticipation in their task. The current study has a more direct measure of error as we used pupil change during the target period as our measure of error.

Using both behavioral evidence and a computational approach, this study sought to pinpoint how infants use prediction error to learn (update prediction) in a systematic way. We examined 8-months-old infants' performance on a modified VExP, where they viewed in left-right alternation, a sequence that is Predictable and Semi-Predictable (**Figure 20**). We measured

both eye movements and pupil size to study prediction and prediction error respectively. If infants' eye movements show a preference for the Semi-Predictable Sequence, we expect to find a better model fit for the PH model, suggesting prediction error taps into infants' attention. On the other hand, if infants' eye movements suggest a preference for the Predictable Sequence, we expect to find a better model fit for the RW model suggesting a more direct role of prediction error in learning.

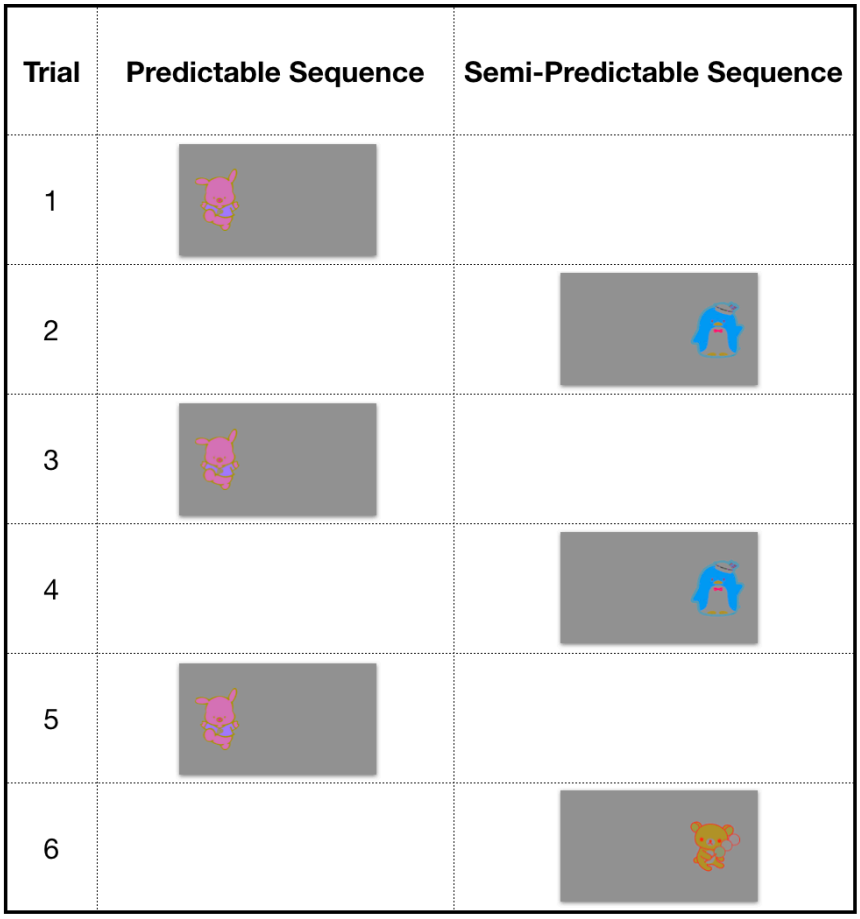


Figure 20. Schematic of the modified Visual Expectation Paradigm during the test phase. Infants saw two types of sequences. The Predictable Sequence (illustrated here on the left) was composed of the Predictable Target (i.e. same target shown repeatedly) and the Semi-Predictable Sequence (illustrated on the right) was composed of the Common Target, shown on 66% of the trials and the Prediction Error Target, shown on the other 33% of the trials. In this example, the experiment begins with the predictable sequence, with the target appearing on the left side of the screen.

Methods

Participants

We included 25 8-months-old (M = 262 days, SD = 7.4 days) infants (12 female, 13 male) in analyses. An additional 20 infants participated but insufficient data (i.e., completed less than 25% of experiment for both sequences) were collected because they moved too much (n = 11), cried (n = 1), or were inattentive (n = 8) and 1 additional infant was excluded because they were outside of the age range and recruited in error. The dropout rate (45.6%) for this experiment is similar to a previous VExP study using almost identical design (Adler & Haith, 2003). Infants in the sample were full-term at birth (37 weeks or greater) with no reported complications. Experiments were approved by Princeton University's Institutional Review Board and informed consent was obtained before beginning the study from the legal guardian of the infant. Families received $10, a t-shirt, and a children's book for their participation.

Stimuli

The visual stimuli (i.e. targets) were three cartoon characters (bunny, bear, pig) and presented 150 pixels to the left or right of the center of the screen. Each target was also paired with an audio stimulus (xylophone, laser sound, magic sound) in order to attract infants' attention to the screen. All targets and background color of the screen were made isoluminant to each other and the room to prevent changes in pupil size as a result of varying brightness. We used an isoluminance script provided by Daniel Swingley (personal communication, David Brainard and Daniel Swingley, 2016) which takes a standard RGB image, corrects the gamma, separates the image's chromaticity and luminosity, and sets the luminance to a constant value, thereby making the image isoluminant, without affecting the chromaticity. We selected a

luminance value of 0.3 to match the lighting in the room and not distort the images and videos to the point of being unrecognizable. Stimuli were presented using MATLAB for Mac (R2015b) and Psychtoolbox (3.0.12 Beta, SVN revision 6399).

Procedure

Participants sat approximately 60 cm from the display monitor and eye tracker (Eyelink 1000, SR Research, Ontario, Canada). The display monitor (Eyelink 1000 eye tracker) was secured to an arm mount facing the participant. The host monitor plus experimenting computer were in front of the experimenter. Before beginning the experiment, a randomized three-point calibration was used. We performed calibration and validation for all participants and only included participants that had good (i.e. the highest level) validation.

Infants sat on their caregivers' laps throughout the experiment. Caregivers were instructed to not interfere with the infant throughout the study and wore a sleeping mask during the experiment, which prevented them from seeing the screen and blinding them to the content of the individual trials (i.e., to prevent biasing of infant behaviour). The experimenter watched the caregiver and infant from the Eyelink 1000 host computer in order to exit the experiment when infants became too inattentive and fussy. Monitoring the host computer also allowed the experimenter to adjust the display monitor as infants moved.

Each trial was composed of a 1200ms delay, followed by 1000ms presentation of the target. The delay allowed infants to make anticipatory eye movements to the location of the upcoming target. Infants completed a total of 108 trials, with the first 12 trials acting as a baseline period during which targets were randomly presented on the left or right side and infants' eye movements and pupil size prior to learning was assessed. Training was important

because it allowed infants to become accustomed to the pace of the experiment and familiar with the location of the stimuli (left and right side of the screen) and to obtain baseline pupil size. The remaining 96 targets constituted the experimental phase during which targets appeared in a left-right alternation. One side of the screen showed the predictable sequence, which was composed of the Predictable Target (i.e. same target shown repeatedly) and the other side of the screen showed the Semi-Predictable Sequence, which was composed of the Common Target, shown on 66% of the trials and the Prediction Error Target, shown on the other 33% of the trial.

Preprocessing

Data Cleaning

Data files exported from the eye tracker were processed to determine pupil change for each participant. First, blinks were identified and linearly interpolated (Addyman, Rocha & Mareschal, 2014; Fawcett, Wesevich & Gredebäck, 2016; Hepach & Westermann, 2013) using the package zoo (Version 1.7-13, Zeileis, Grothendieck, Ryan & Andrews) for the R software (Version 3.2.1; R Development Core Team, 2015). Next, we removed data collected when infants were gazing off-screen (i.e. X and Y gaze coordinates were outside the bounds of the screen dimension, 1280 x 1024). Lastly, we excluded infants who completed less than 25% of the trials for each sequence (Rose et al., 2002).

Eye Movements

Gaze coordinates were coded as looking at upcoming/current target location, center of the screen, previous target location or missing. Following previous work (Canfield & Haith, 1991; Haith et al., 1988; Jacobson et al., 1992; Rose et al., 2002; Wentworth & Haith, 1992; Wentworth, Haith & Hood, 2002), eye movements were considered to be anticipatory (AEMs) if

infants looked to the upcoming/current target location between 0-1400ms of a trial or from the beginning of the trial to the first 200ms of target onset (**Figure 21**). The VExP literature is fairly inconsistent in terms of anticipatory cut-off, with 133ms, 167ms, and 200ms being the most common cut-offs. Reznick, Chawarska & Betts (2000) found that when categorizing AEMs using different cut-offs, specifically 133ms and 200ms after target onset, it did not make a difference in their results. Eye movements that occurred after 1400ms were considered to be reactive eye movements (REMs). Reaction time (RT) to target for the REMs was determined by subtracting when the REM occurred from when the target appeared. We calculated proportion of AEMs and REM RT during the test phase separately for each sequence (Predictable and Semi-Predictable) and used the first AEM and REM for analysis if infants made multiple eye movements during a single trial (Romberg & Saffran, 2013).

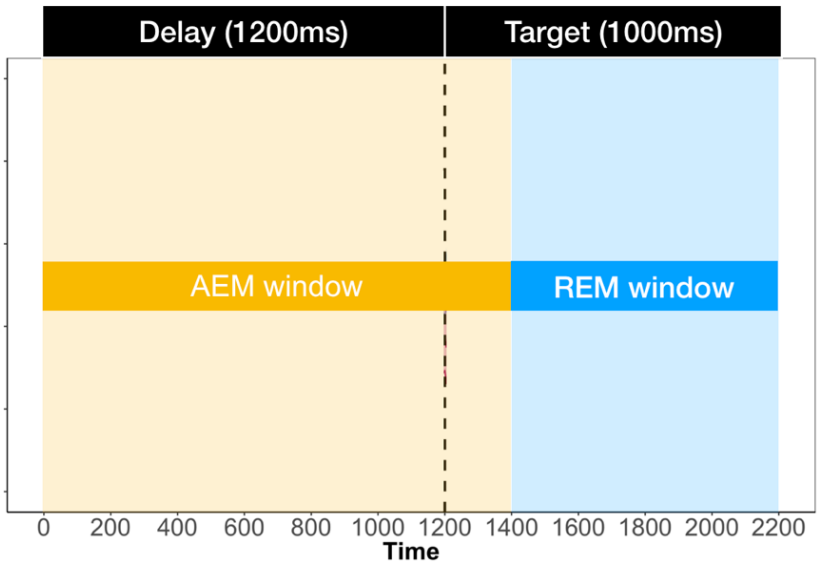


Figure 21. Schematic of trial. The time-course of each trial was identical for both predictable and semi-predictable sequence. Each trial began with delay for 1200 ms, followed by the target for 1000 ms. On each trial, we categorized eye movements into AEMs (orange) if infants first looked to the target during 0 -1400 ms, and REM (blue) if it occurred after 1400 ms. REM RT was determined by subtracting time when target appeared (1200ms) from when infants initiated the REM.

Pupil Change

We analyzed pupil size during the test phase by calculating percentage change from baseline. We had two baseline measurements, one for the left and one for the right side of the screen, as pupil size is known to be distorted differently when looking away from the center of the screen (Hayes & Petrov, 2016). In order to get a precise measurement of baseline pupil size, pupil size was measured during the training phase only when the target was on the screen and infants were looking at the target. The appropriate baseline measure was then subtracted from the pupil diameter measurement, divided by baseline, multiplied by one hundred, to get percentage change from baseline for each trial. We also excluded trials with average pupil change larger than 2.5 standard deviations from the mean for each subject during the target period for each subject, which is likely a result of eye movement.

Learning Models

We used two error-driven learning models, the RW model and PH model, to examine the trial-by-trial changes of prediction and prediction error in the task. We operationalized prediction error as the magnitude of pupil change during the target period for each trial (Kloosterman et al., 2015; Lavin, San Martín & Rosales Jubal, 2014; O'Reilly et al., 2013; Preuschoff, Hart & Einhauser, 2011; Sirois & Jackson, 2012, Zhang et al., 2018). For each participant, we fitted a generative implementation of each learning model to their PDR response.

For the RW model, prediction error is calculated for each trial, t:

$$\delta(t) = O(t) - P(t)$$

where: $O(t)$ is the appearance of the Common Target for the Semi-Predictable Sequence and appearance of the Predictable Target for the Predictable Sequence. $P(t)$ is the predicted probability of that appearance. It is important to note that in the Semi-Predictable Sequence, the

Common Target is the target that infants should be predicting to see because it appears more often. Thus, the Common Target is coded as a 1 in the learning model and the Prediction Error Target is coded as a 0 because it is error inducing.

In addition, predictions are updated for each trial based on the prediction error and learning rate, where α is the learning rate:

$$P(t+1) = P(t) + \alpha\delta(t)$$

For the PH model, prediction error is calculated in the same way as the RW model, for each trial, t:

$$\delta(t) = O(t) - P(t)$$

Next, predictions are updated for each trial based on the prediction error and learning rate (i.e. attention), where κ is a scaling factor:

$$P(t+1) = P(t) + \kappa\alpha(t)\delta(t)$$

Lastly, the learning rate (α) is updated for each trial based on the prediction error squared and learning rate, where ε is the rate the learning rate updates:

$$\alpha(t+1) = \alpha(t) + \varepsilon\left(\left(\delta(t)\right)^2 - \alpha(t)\right)$$

First, using all the trials, we obtained the initial learning rate, and initial prediction for the RW model and PH model. Next, we took these initial parameters, and used them as the initial parameters and modeled the two sequences (predictable and semi-predictable) separately for each model. When modelling the sequences separately, the initials parameters are different for each sequence in order to minimize mean squared error (MSE). Hence, we used the initial predictions obtained from the modelling of all trials because infants should have the same initial predictions for both sequences. We decided to measure the sequences separately because we

hypothesized differences between the learning trajectories and modelling the sequences together would not allow such a comparison.

Technical Explanation Of Model Fitting Process:

For both models, we fitted the individual models by minimizing the Mean Square Error (MSE) between the PDR and δ^2 (level of surprise or magnitude of prediction error regardless of the direction of the error) throughout the experiment for each participant. The parameters that were used to fit the model to the PDR are the initial prediction, $P(0)$ for the appearance of the target and the initial learning rate, $\alpha(0)$. These parameters were optimized using the Nelder-Mead simplex method (Lagarias et al., 1998) when modelling all the trials and were predefined when modelling the sequences separately. The parameters were restricted to be non-negative and to be not greater than 1. For each trial of every subject, the fitted model estimated the prediction error, $\delta(t)$, and prediction, $P(t)$. For each participant, we initialized the fit from evenly spaced starting points and used the lowest achieved MSE as the best fit.

Results

The current study integrated eye movements, as measures of infant prediction and learning, and pupillometry, as a measure of prediction error, with computational learning models to determine the relationship between prediction error and learning in infancy. First, we examined the behavioral data, specifically how AEMs, REMs and pupil change differed between the two sequences (Predictable and Semi-Predictable). This allowed us to evaluate whether these measures replicated previous findings and reflect cognitive differences at a broad level. For null results, we followed up with exploratory analyses and analyzed how the measures differed between target (Predictable Target from the Predictable Sequence, Prediction Error Target and

Common Target from the Semi-Predictable Sequence).

Anticipatory Eye Movements (AEMs) By Sequence

Previous work found that infants generated more AEMs for the Predictable Sequence than the Unpredictable Sequence (Adler & Haith, 2003; Wentworth & Haith, 1992). Although we replaced the Unpredictable Sequence with a Semi-Predictable Sequence, we conducted the same analysis to determine if infants anticipated more for one sequence compared to the other. We calculated the proportion of AEMs for each infant as the number of correct AEM divided by the total number of valid trials (Adler & Haith, 2003). On average, infants contributed 9.2 (*range* = 2 – 19, *SD* = 5.24) number of correct AEM trials and 42.58 (*range* = 36 – 48, *SD* = 3.37) number of valid trials for the Predictable Sequence and 9.8 (*range* = 3 – 23, *SD* = 4.63) number of correct AEM trials and 42.33 (*range* = 36 – 48, *SD* = 3.16) number of valid trials for the Semi-Predictable Sequence.

A paired t-test was performed to compare the proportion of AEMs made for the Predictable Sequence (M = 21.3%, SE = 2.3%) and the Semi-Predictable Sequence (M = 23.5%, SE = 2.4%) (**Figure 22**). Unlike previous studies (Adler & Haith, 2003; Wentworth & Haith, 1992), our results showed no difference between the two sequences ($t(23) = -1.55$, $p = 0.55$). Comparing our numbers with previous findings, the average proportion of AEM for the Predictable Sequence is similar to what was found in Adler and Haith (2003). However, infants generated more AEMs for the Semi-Predictable Sequence compared to only 11% of AEMs for the Unpredictable Sequence, across multiple studies, in Adler and Haith (2003). Thus, it is possible that because we replaced a completely unpredictable sequence with a semi-predictable sequence, the differences between the two sequences in the experiment became less obvious, and

consequently, it became more difficult for infants to differentiate the two sequences. We will return to these findings in the discussion.

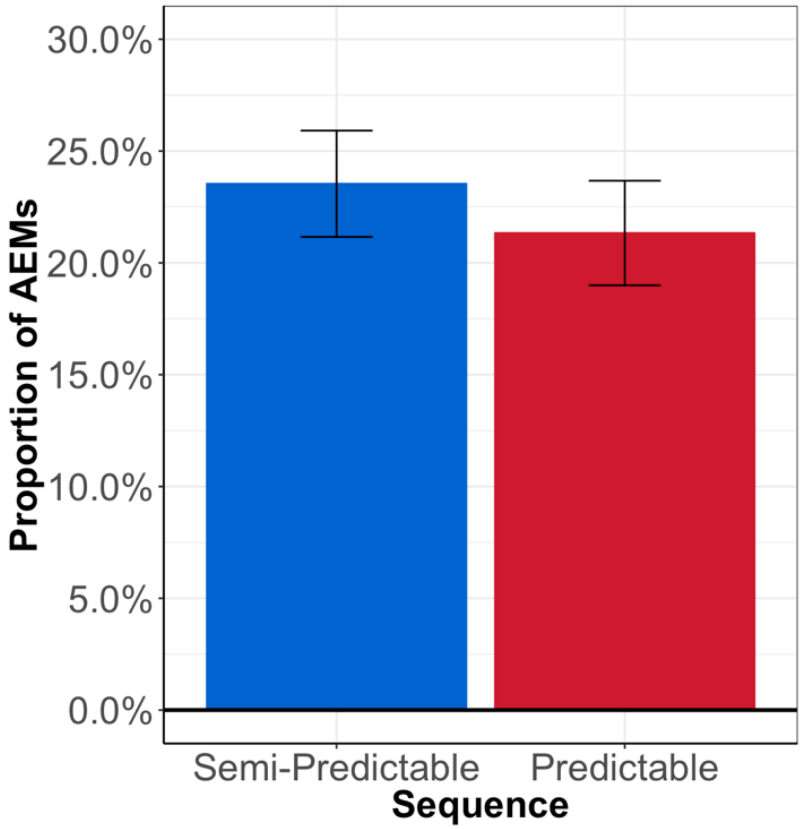


Figure 22. Proportion of AEMs (the number of correct AEMs / total number of valid trials) as a function of sequence (Semi-Predictable and Predictable). Vertical error bars indicate +/– SE.

Reactive Eye Movements (REMs) By Sequence

It is possible that differences between the two sequences are reflected in the REMs despite no significant findings with the AEMs, as studies have found differences in one measure but not the other (Adler & Haith, 2003; Wentworth & Haith, 1992). On the trials during which infants did not execute an AEM, the REM RT was calculated by subtracting when the target appeared (1200ms) from when the REM was initiated. Next, we calculated the median REM RT for each participant, instead of the mean, because infants' eye movement latencies tend to be

skewed (Adler & Haith, 2003; Canfield & Haith 1991; Wong-Kee-You & Adler, 2016). The mean RT for the Predictable Sequence was 476.75 ms (*range* = 304 ms – 668 ms, *SD* = 71.31 ms) and Semi-Predictable Sequence was 498.56 ms (*range* = 396 ms – 652 ms , *SD* = 82.48 ms). On average, infants contributed 23.54 (*range* = 12 – 40, *SD* = 8.2) number of REM trials for the Predictable Sequence and 23.24 (*range* = 9 – 41, *SD* = 8.34) number of REM trials for the Semi-Predictable Sequence.

First, both sequences' REM RT were significantly faster for the test phase compared to training [semi-predictable: $t(23) = 4.3$, $p < 0.001$; predictable: $t(23) = 7.08$, $p < 0.001$] (**Figure 23**). There are mixed results with respect to how REM RT during test phase compare to REM RT during baseline. Adler and Haith (2003), did not find a difference in REM RT between test and baseline, however, Jacobson et al. (1992) did and attributed this faster REM RT during test as a result of facilitation from the alternating pattern. In addition, the REM RT of the two sequences during test phase did not differ from each other ($t(23) = 1.05$, $p = 0.3$). Similar to Adler and Haith (2003), they did not find a difference in REM between the Predictable and Unpredictable sequences, and they used this null finding to dissociate AEM from REM because they found a difference in proportion of AEMs. These results suggest that although infants did elicit faster REMs for test phase compared to training, there are still no behavioral differences between the two sequences.

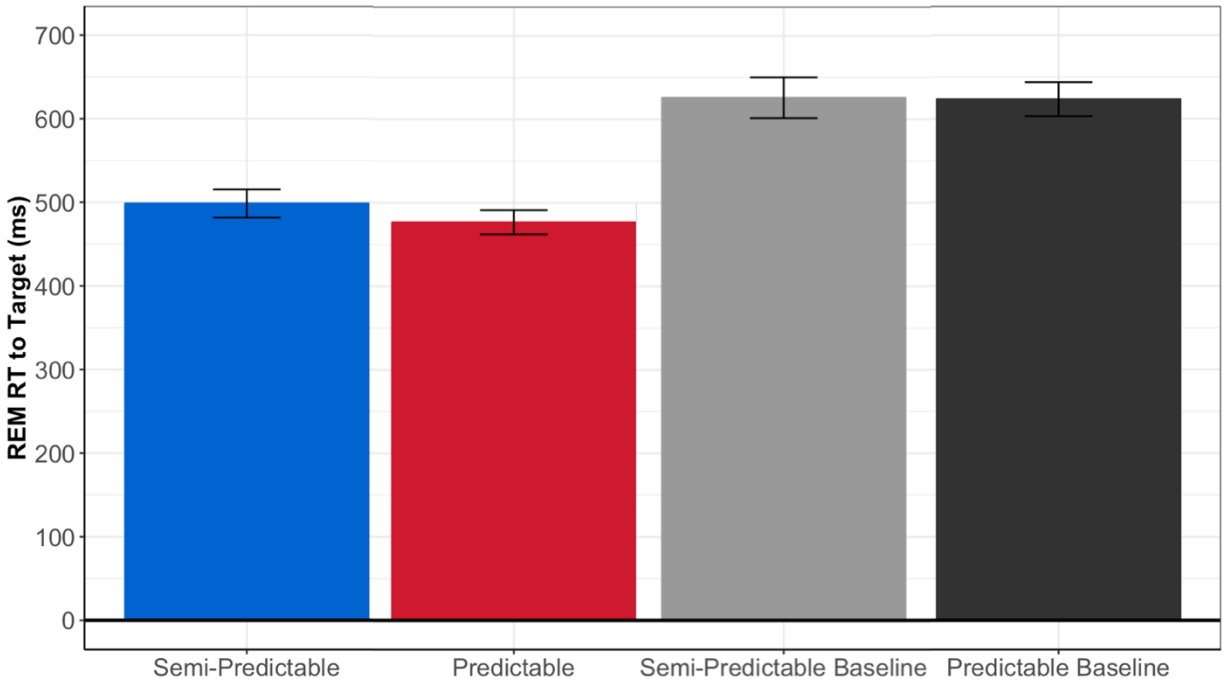


Figure 23. Median REM RT as a function of sequence (Semi-Predictable and Predictable) and its respective baseline. Vertical error bars indicate +/– SE.

Reactive Eye Movements (REMs) By Target

Although REM RT did not differ between sequences, we conducted an exploratory analyses to determine if REM RT would differ between targets. We were specifically interested in whether the REM RT for the Prediction Error Target would be different than the other two targets, because it is possible that behavioral differences are constrained to only the Prediction Error Target, which occurs 33% of the time in the Semi-Predictable Sequence. Thus, combining the Prediction Error Target and Common Target during analyses that contrast the two Sequences might eliminate the effects of prediction error. On average, infants contributed 23.54 (*range* = 12 – 40, *SD* = 8.2) number of trials for the Predictable Target, 15.88 (*range* = 7 – 28, *SD* = 5.6) number of trials for the Common Target and 7.36 (*range* = 1 – 14, *SD* = 3.2) number of trials for the Prediction Error Target.

A repeated-measures one-way analysis of variance (ANOVA) was performed to compare the median reaction time of REMs for the three targets (Predictable Target, Common Target and Prediction Error Target) (**Figure 24**). The main effect of target was significant, $F(2, 69) = 3.6$, $p = 0.03$, indicating that infants did differ in REM RTs. Interestingly, pair-wise comparisons (Tukey's HSD) showed infants made significantly slower REMs for Prediction Error Target compared to Predictable Target (65 ms, $p = 0.044$) and slower REMs for Prediction Error Target compared to Common Target was approaching significance (57.5 ms, $p = 0.084$). This was an unexpected finding as we hypothesized that prediction error might capture infants' attention allowing them to orient to the target faster. However, it is possible that predictability facilitated the REM RTs for the Predictable Target and prediction error generated by an unexpected target, slowed the REM RTs for the Prediction Error Target.

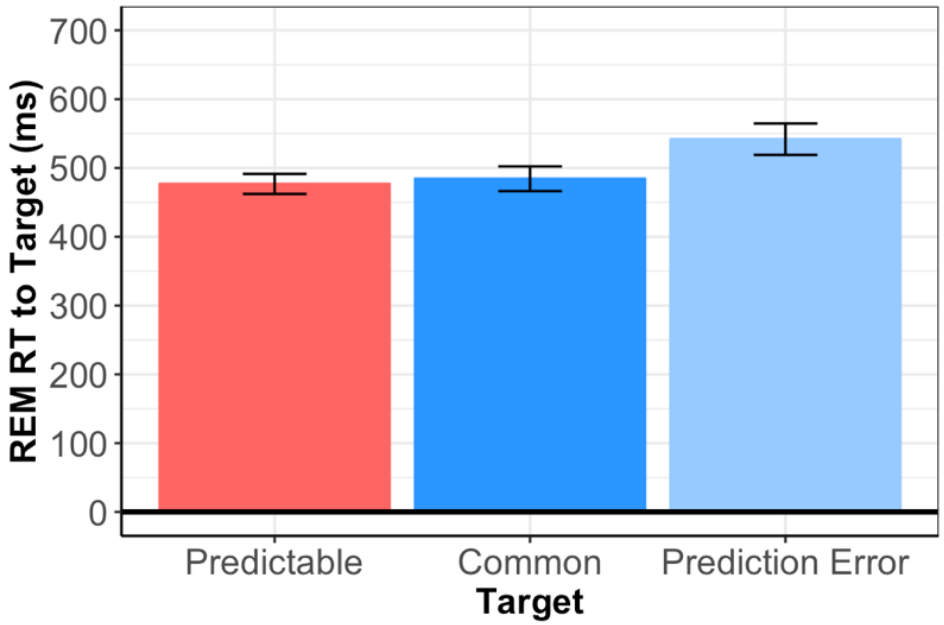


Figure 24. Median REM RT as a function of target (Predictable Target from the Predictable Sequence, and Common Target and Prediction Error Target from the Semi-Predictable Sequence). Vertical error bars indicate +/– SE.

Pupil Change By Sequence

After calculating pupil change from baseline for every single data point, we computed average pupil change for when infants were looking at the target during the target-on period, for every trial. For this analysis, infants contributed an average of 31.25% (*range* = 20 – 44, *SD* = 7.71) number of trials for the Predictable Sequence and 30.75% (*range* = 16 – 44, *SD* = 6.95) number of trials for the Semi-Predictable Sequence. We expected to find a larger average pupil change for the Semi-Predictable Sequence compared to the Predictable Sequence because previous studies have shown that larger pupil change is indicative of prediction error. However, our results revealed that the two sequences did not have significantly different pupil change ($t(23) = 0.98$, $p = 0.34$, **Figure 25**). This was an unexpected finding given the numerous previous studies establishing a relationship between larger pupil change and prediction in both adults and infants.

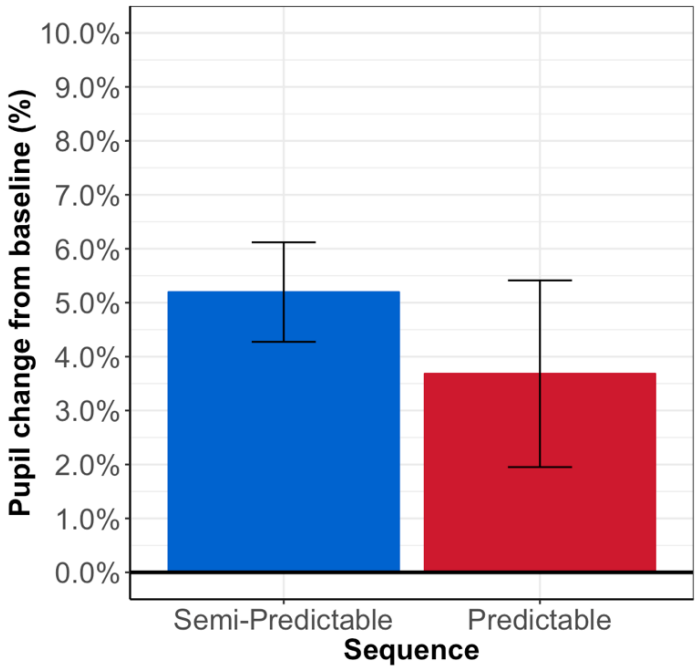


Figure 25. Average pupil change during when infants are looking at the target on the screen (1200-2200ms), grouped by sequence (Predictable and Semi-Predictable). Vertical error bars indicate +/– SE.

Pupil Change By Target

We conducted an exploratory analyses to investigate whether pupil change reflected prediction error on a target level, instead of sequence level. It is possible that averaging the pupil change for the Prediction Error Target and Common Target together for the Semi-Predictable Sequence, washed out the effects of the Prediction Error Target. Infants contributed an average of 31.25 (*range* = 20 – 44, *SD* = 7.71) trials for the Predictable Target, 20.64 (*range* = 10 – 30, *SD* = 5) number of trials for the Common Target and 9.8 (*range* = 5 – 14, *SD* = 2.56) trials for the Prediction Error Target. However, a repeated-measures one-way analysis of variance (ANOVA) was performed and we determined that pupil change was not significantly different across the three targets ($F(2, 69) = 1.5$, $p = 0.23$, **Figure 26**). Overall, although pupil change is trending in the right direction (i.e. it's larger for the Prediction Error Target and Semi-Predictable Sequence), it does not appear to significantly reflect prediction error in this task.

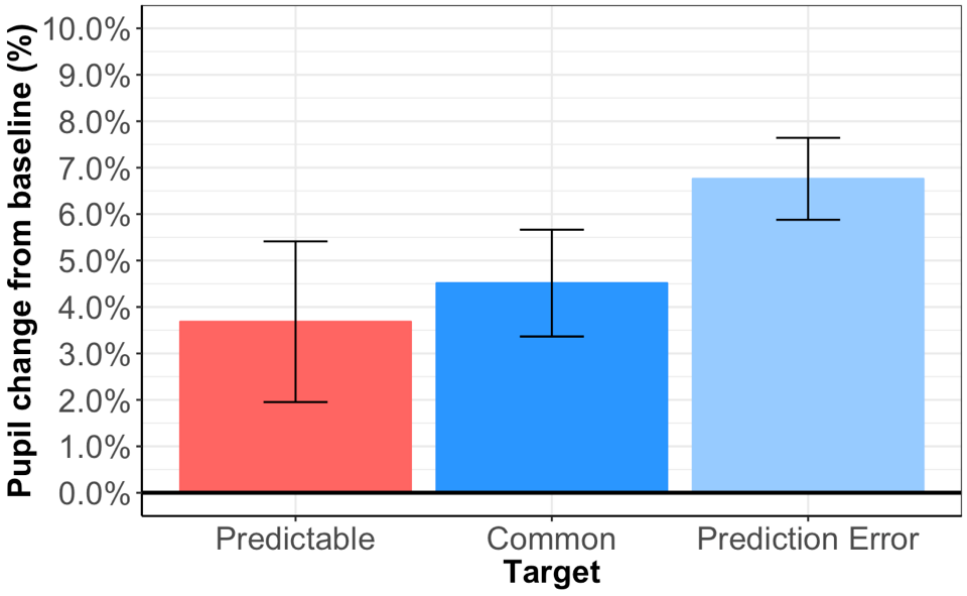


Figure 26. Average pupil change during when infants are looking at the target on the screen (1200-2200ms), grouped by target (Predictable Target from the Predictable Sequence, and Common Target and Prediction Error Target from the Semi-Predictable Sequence). Vertical error bars indicate +/– SE.

Learning Models

We incorporated the RW model and PH model into our analyses as a way to systematically study prediction error in infants. Although we did not find a significant relationship between pupil change and prediction error using behavioral data, it is possible that the combination of pupil change and error driven learning models can still reveal meaningful information about how infants learn. First, we fitted each model, RW and PH, using the average pupil change during the target period for every trial. The RW model revealed infants' initial prediction, $P(0)$ ($M = 0.48$, $SD = 0.31$) and learning rate $\alpha(0)$ ($M = 0.052$, $SD = 0.12$, **Figure 27A**), whereas the PH model revealed infants' initial prediction $P(0)$ ($M = 0.5$, $SD = 0.29$), attention $\alpha(0)$ ($M = 0.27$, $SD = 0.41$), and rate at which they update their attention, ε ($M = 0.34$, $SD = 0.44$, **Figure 27B**).

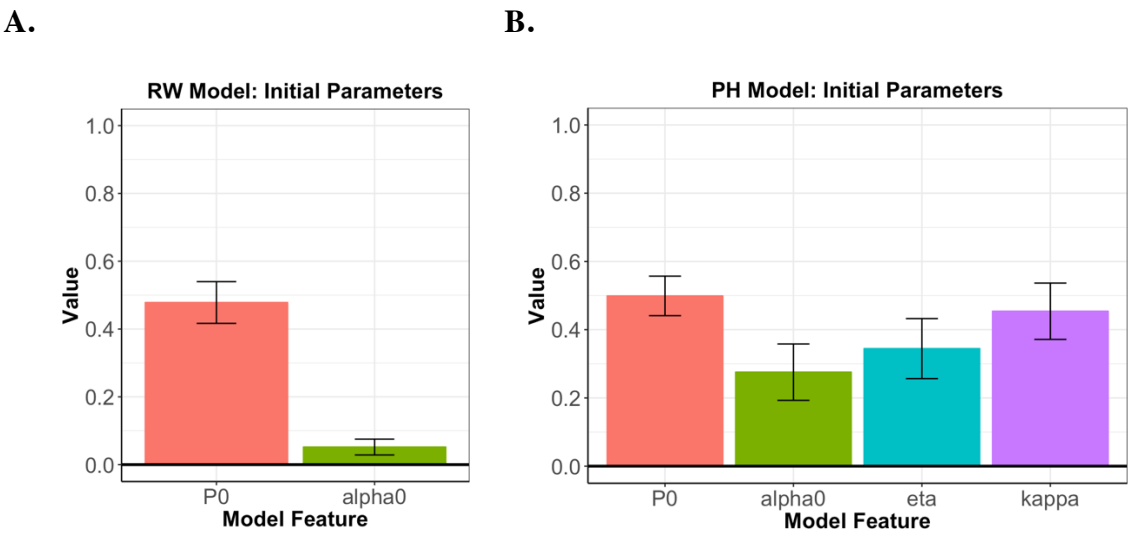


Figure 27. Initial Parameters for the (A) RW Model and (B) PH Model. Vertical error bars indicate +/– SE.

Next, we used the initial parameters generated by the models to calculate trial-by-trial changes in Prediction (P) and Prediction Error (δ^2) for both models. We decided to conduct the

trial-by-trial analyses separately for the two sequences (e.g. run trial by trial analyses of the RW model using only data from the Predictable Sequence and another RW model using only data from the Semi-Predictable Sequence), because according to previous VExP studies, infants exhibited different behaviors for the two sequences (Adler & Haith, 2003; Jacobson et al., 1992; Wentworth & Haith, 1992). Thus, it is possible that infants built two different learning models, one for each sequence.

Figure 28 shows the trial by trial changes in Prediction and Prediction Error for the two sequences separated for the two models. Before interpreting the graphs, we wanted to confirm that the Predictions generated by the models reached the appropriate value (1.00 for the Predictable Sequence since the Predictable Target always appeared, and 0.66 for the Semi-Predictable Sequence since the target was the Common Target). However, it is immediately clear that Prediction for the Predictable Sequence did not reach 1 for either learning model, suggesting that these models may not be appropriately reflecting the current task.

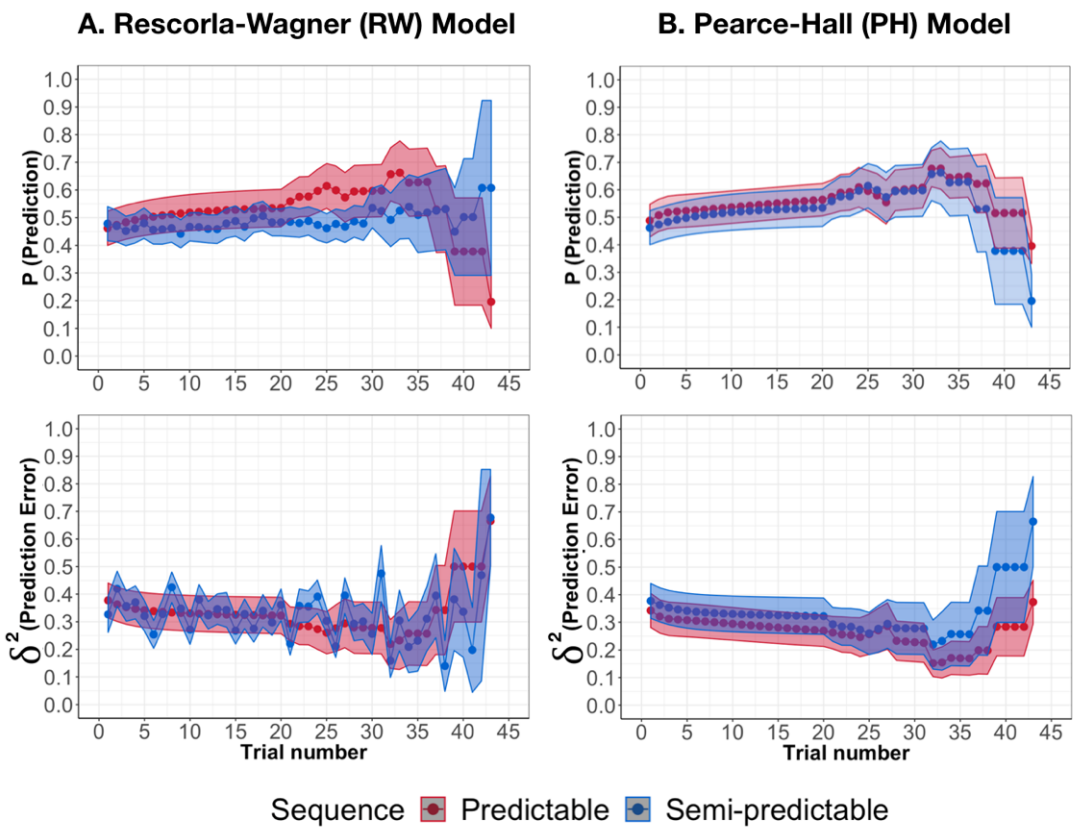


Figure 28. Trial-by-trial changes in model parameters, Prediction and Prediction Error, generated separately for each sequence for the (A) RW Model (B) PH Model.

To confirm that the failure of the models to reflect the structure of the task is not due to an issue with the coding of the model itself, we created an ideal model that did not require the pupil responses. Instead, we set model's initial parameters to ideal values such that the learning models should definitely converge to the correct Prediction for both sequences. The ideal initial parameters we used for the RW models were $P(0) = 0.5$, $\alpha(0) = 0.2$, and PH model were $P(0) = 0.5$, $\alpha(0) = 0.2$, $\varepsilon = 0.2$, $\kappa = 1$. Because the initial prediction, P(0), was set to 0.5, and the initial learning rate $\alpha(0)$ and the rate at which you update your learning rate (ε) were both set to 0.2, prediction (P) on a trial-by-trial basis will be updated by a significant amount after exposure to just a single trial. Thus, we can be confident that the model will learn to update Prediction to the

correct value. **Figure 29** confirms that the ideal learning models are performing appropriately (Prediction for the Predictable Sequence reaches 1 for both models) when given the ideal parameters, suggesting the code is accurate and the issue behind **Figure 28** is due to a larger issue of using the wrong learning models to fit the infant pupil data from the task.

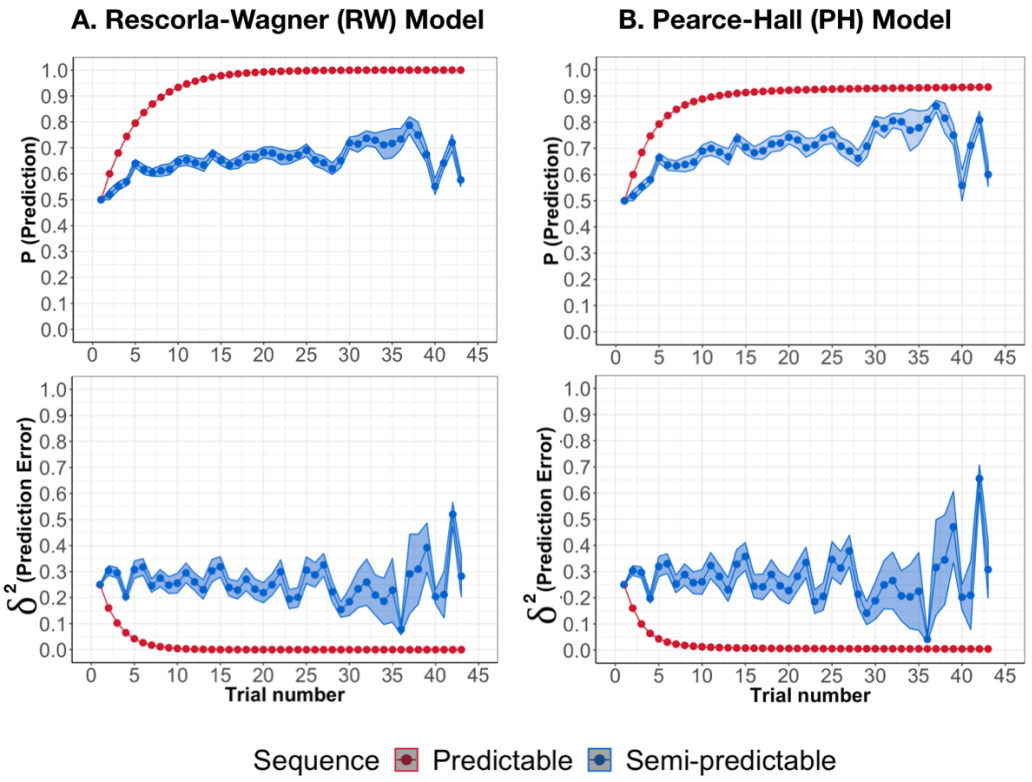


Figure 29. Trial by trial changes in model parameters, Prediction and Prediction Error, generated separately for each sequence using ideal initial parameter values for the (A) RW Model (B) PH Model.

Having confirmed that under ideal circumstances these models can reflect the structure of the task, we turned to evaluate whether pupil responses reflect prediction error, the parameters that they are assuming in the computational learning models. In other words, it may be that forcing the average pupil change to operate as a prediction error parameter in these learning models is not appropriate and is leading to convoluted results. To explore this possibility, we

determined the relationship between Prediction Error (δ^2) from the ideal learning models and pupil responses collected from the infants. Specifically, the trial-by-trial Prediction Error (δ^2) from the ideal learning model was used as the independent variable and infants' average pupil change during target period was the dependent variable in a linear regression model. We conducted four linear regression models, one for each unique pair of learning model (RW and PH) and sequence (Predictable and Semi-Predictable). If Prediction Error (δ^2) in the learning model truly maps onto pupil change in the task, we expect to see a significantly positive coefficient for both the RW and PH model but only in the Semi-Predictable Sequence, since that's where prediction error is generated. However, the regressions revealed that Prediction Error (δ^2) in the learning model did not predict pupil change in the task for the Semi-Predictable Sequence for either model (RW model: $\beta = -0.005, p = 0.86$; PH model: $\beta = -0.001, p = 0.96$). In addition, we found that Prediction Error (δ^2) in both learning models negatively predicted pupil change in the Predictable Sequence (RW model: $\beta = -0.16, p = 0.030$; PH model: $\beta = -0.17, p = 0.033$). In other words, the larger the Prediction Error (δ^2) as determined by the ideal models, the smaller the pupil change in the Predictable Sequence during the task. This result is hard to interpret both because the Predictable Sequence was designed to elicit no prediction error and that this is opposite of the hypothesized direction. Taken together, the Prediction Error (δ^2) parameter in the RW model and PH model did not map onto pupil change in the task. This is likely arising from the fact that we found no pupil change difference found between the two sequences, despite the fact that one sequence was designed to generate no prediction error and one was designed to generate prediction error.

Discussion

The goal of this study was to investigate prediction error as a possible mechanism underlying infant learning. We used a modified the visual expectation paradigm (VExP) to examine how 14-month-old infants' eye movements (AEMs, REMs) and pupil size differed when viewing a Predictable Sequence, designed to generate no prediction errors versus Semi-Prediction Sequence, designed to generate prediction errors. We hypothesized that learning and preference for a sequence will be manifested in the form of higher proportion of AEMs and faster REMs. In addition, we hypothesized that their experience of prediction error, measured using pupil change, would predict learning. Lastly, we sought to use this study to differentiate how infants use prediction error: is it more accurately described by the RW model or PH model?

Focusing on eye movements, we did not find differences in the proportion of AEMs or speed of REMs between the Predictable and Semi-Predictable Sequence. We believe these null findings are a result of replacing the Unpredictable Sequence used in the original VExP studies with the Semi-Predictable Sequence because it diminished the contrast between the two sequences used in the study. In other words, infants had a more difficult time separating the Semi-Predictable Sequence from the Predictable Sequence, resulting in similar eye movements for both sequences.

We also explored the possibility that REM RTs might differ for each target type (Predictable Target, Common Target and Prediction Error Target) used in the task and comparing REM RT by sequence required us to group the Common Target and Prediction Error Target together. When we compared REM RTs for each target type, we found that infants generated faster REM RTs for the Predictable Image and Common Image compared to the Prediction Error Target. This was a surprising finding given our hypothesis that prediction error

is beneficial for learning, which should result in the fastest REM RT for the Prediction Error Target. However, it appears in this particular paradigm, prediction error slows down REM RTs, and predictability speeds it up. This aligns with the results of previous VExP studies showing infants have a preference for the Predictable Sequence over Unpredictable Sequence (Adler & Haith 2003; Jacobson et al., 1992; Wentworth & Haith, 1992). One hypothesis behind a preference for predictable targets could be the left-right target alternation in VExP studies combined with a predictable target puts the task difficulty in a sweet spot (i.e. the perfect amount of difficulty for infants to be really engaged and learn) and viewing unpredictable images pushes the task into too difficult such that it hinders infants learning (Kidd et al., 2012; 2014). In our study, this hinderance manifested itself as slower REM RT to the Prediction Error Image.

Moving onto pupil change, we hypothesized that infants' pupil change would be larger for the Semi-Predictable Sequence compared to Predictable Sequence, as prediction error has been shown to be reflected in larger pupil change. However, we found no differences in average pupil change during the Target period for the two sequences and there are a few reasons behind why this is the case. Starting with the most technical reason, Hayes & Petrov (2015) demonstrated that when using a stationary eye-tracker mounted below the computer screen, pupil size is most accurate when participants are staring in the center of the screen. When participants move their gaze to different locations on the screen, the shape of the recorded pupil changes from a circle to an ellipse (Atchison & Smith, 2000). Consequently, this change in the recorded geometry of the pupil is accompanied by a change in the measured pupil size. Specifically, they found that the pupil size of the artificial eye shrunk as it moved to the left and enlarged as it moved to the right. The change in pupil size due to gaze position is problematic because the essence of VExP is in its use of the left-right alternating sequence to elicit predictive eye

movements. However, other infant studies that have combined eye movements and pupil change (Cheng, Kaldy & Blaser, 2019; Kleberg, del Bianco & Falck-Ytter, 2019; Ross-Sheehy & Eschman, 2019) have found differences in pupil change in their study, suggesting it is not the limiting factor. Furthermore, we carefully thought about and designed an appropriate baseline for this task, so pupil distortion as a result of eye movement shouldn't be the issue per se.

Another possibility for our lack of pupil change finding is our Target period was not long enough for the pupil change to fully respond and return to baseline. Studies with adults have shown that the pupil response is a fairly slow response and takes approximately 6 seconds to return to baseline (Murphy, Van Moort, Nieuwenhuis, 2016). In the successful infant studies that combined eye movements and pupil change, the period during which they measured pupil change was 3000 ms compared to our 1000 ms (Cheng et al., 2019; Kleberg et al., 2019; Ross-Sheehy & Eschman, 2019). We did not choose a longer target period because classic VExP studies have a target period of 700 – 1000ms (Adler & Haith, 2003; Canfield & Haith, 1991; Haith et al., 1988; Jacobson et al., 1992; Reznick et al., 2000; Wentworth & Haith, 1992) and we did not want to alter the paradigm more than we had to.

The last, and more likely reason for our lack of pupil change finding is our replacement of the Unpredictable Sequence, used in classic VExP studies (Adler & Haith, 2003; Wentworth & Haith, 1992), with the Semi-Predictable Sequence. We believe this is the more likely reason because we found that the average proportion of AEM for the Predictable Sequence was similar to previous study (Adler & Haith, 2003) but the proportion of AEM for the Semi-Predictable Sequence was much higher than typically found in the Unpredictable Sequence. These results suggest that infants are learning and performing the task properly, as evidenced by the proportion

of AEM for the Predictable Sequence. However, infants performed equally well during the Semi-Predictable Sequence, resulting in no differences between the two sequences.

Moving onto the modelling results, we used average pupil change during the target period to fit two well-established error driven learning models: the RW model and PH model. We did not find any differences in the parameter and trajectory outputs of the RW and PH model. Given our lack of pupil results, it may not be surprising that our modelling approach also revealed no differences. However, we know that pupil change can be effectively used to dissociate the two learning models (Koenig et al., 2018). Thus, future studies should continue to pursue this line of research in infants because the answer will help us more precisely determine how infants are using prediction error to learn.

Future studies should also continue to incorporate eye movements and pupil change in the same task, despite the issues that were present in the current study. First, we believe that there are steps researchers can take to minimize the pupil distortion caused by eye movements such as: carefully selecting a baseline, making stimuli isoluminant and counterbalancing. Second, the timing of the current study might have contributed to the lack of pupil change and we advise future infant pupillometry studies to have a sufficient period of at least 3000ms to measure pupil size. As mentioned briefly, three studies have successfully integrated eye movements and pupil change in the same task and revealed new insights about how infants process information on a moment by moment basis and how that processing relate to different cognitive abilities. For example, Cheng et al. (2019) measured eye movements and pupil change during a visual working memory task and found that infants with larger pupil change during encoding remembered more object-location bindings during test, based on anticipatory eye movements. In addition, Kleberg et al. (2019) measured eye movements and pupil change during a visual search

task and found that 6 months old infants had the best performance (i.e. highest likelihood of target fixation) when pupil change was considered to be at an intermediate level rather than too small or too large. Ultimately, the approach of combining eye movements, pupillometry and learning models will help provide a deeper understanding of how infants learn, allowing researchers to focus on trajectory and change and pinpoint mechanisms.

Chapter V: Is prediction error the mechanism behind infants' ability to update anticipation?

Prediction is the general ability to use current knowledge to make expectations about the future. This ability is believed to support learning by connecting what has been learned in the past with the current environment. Although prediction is often studied as a single process, prediction has been shown to be composed of several different phenomena (Bubic, von Cramon, & Schubotz, 2010). First, **anticipation** is the process of formulating an **expectation**, or a representation of what you think will happen in the near future (LaBerge, 1995). Next, the expectation and sensory information from the real world are compared. If there is a mismatch between the expectation and sensory information, a **prediction error** occurs (Friston, 2005; Hollerman & Schultz, 1998; McClure, Berns, & Montague, 2003; Rao & Ballard, 1999). Prediction errors triggers **updating**, a process that involves learning from past experiences and fine tuning of expectations to be more accurate in the future (O'Reilly et al., 2013). Breaking down prediction into its component parts allows for a detailed investigation of the mechanisms involved in prediction (for a review on how expectations shape perception, see de Lange, Heilbron, and Kok, 2018; Summerfield & Egner, 2009; for a review on the impact of prediction error see den Ouden, Kok, de Lange, 2012). Overall, research on prediction and it's subprocesses has established predictive processing as having a widespread influence on how we process and respond to sensory information (Bar, 2007; Bubic et al. 2010; Koelsch, Vuust & Friston, 2019; Misyak, Christiansen & Tomblin, 2010).

Prediction also has a long history in developmental research. Studies show evidence for some aspects of predictive processing very early in life (Aslin, 1981; Canfield & Haith, 1991) and across various domains such as pro-social behavior (Jin, Jouston, Baillargeon, Groh &

Roisman, 2018), action (Falck-Ytter et al., 2006; Gredebäck, Lindskog, Juvrud, Green & Marciszko, 2018), language (Lew-Williams & Fernald, 2007; Ylinen, Bosseler, Junttila & Huotilainen, 2016), and vision (Rose, Feldman, Jankowski & Caro, 2002).

Considering the subcomponents of prediction, anticipation has been examined using anticipatory eye movements (AEMs), or eye movements to a location before the target appears. Beginning with classic Visual Expectation Paradigms (VExP) created by Marshall Haith and colleagues (Canfield & Haith, 1991; Haith, Hazan, & Goodman, 1988), these studies revealed that infants as young as 3 months old are capable of anticipating upcoming images when the spatial location follows a simple pattern (Canfield & Haith, 1991; Haith, Hazan, & Goodman, 1988). Specifically, infants made AEMs to the location of the image, suggesting infants are capable of engaging in predictive processing (for a review, see Wentworth, 2009).

Prediction error (i.e. surprise) in developmental research is classically investigated using violation of expectation (VOE) paradigms and looking time (Baillargeon, 1987; Jackson & Sirois, 2009; Sirois & Jackson, 2012). In a seminal study by Stahl and Feigenson (2015), they found that 11-month-old infants who viewed a video that violated their expectation of how a toy should behave (e.g. a floating car), resulting in a prediction error, were better at learning information about the toy compared to infants who viewed a video that followed their expectation (e.g. a non-floating car). Stahl and Feigenson (2015) was the first study to establish a connection between prediction error and learning in infants, a topic that has since gained traction. For example, Ylinen et al. (2016) found that 12-month-olds' electrophysiological brain responses to novel word forms (interpreted as a prediction error response) correlated with their vocabulary score. These studies provide some initial evidence for the importance of prediction error for early learning, however, a robust connection has yet to be established.

Out of all the components of prediction, prediction error is arguably the most relevant for learning, and thus, is particularly important for developmental research and is our focus for the current paper. In fact, there are well-validated learning models, that explicitly establish a relationship between prediction error and learning. The Rescorla-Wagner (RW) model (Rescorla & Wagner, 1972) is a classic associative learning model that states prediction errors will result in corresponding changes in associations (i.e. learning) and no change will occur if the error is zero. Another classic learning model, the Pearce and Hall model, states the amount of attention allocated to a cue, which improves learning for the cue, is determined by the prediction error (Pearce & Hall, 1980; Pearce, Kaye, & Hall, 1982; Pearce & Mackintosh, 2010). Furthermore, in this model, uncertain cues that elicit more prediction error will attract more attention than the consistent cues. Taken together, these models formalize the view that prediction errors create the opportunity to update one's anticipation and expectation or associations based on new information. These models create a direct link between prediction error and learning.

Recently, pupillometry, or the measurement of pupil size, has become a popular method for measuring prediction error for both adults (Koenig, Uengoer, & Lachnit, 2018; Nassar et al., 2012; O'Reilly et al., 2013; Zhang, Jaffe-Dax, Wilson, & Emberson, 2018) and infants (Gredebäck & Melinder, 2010; Jackson & Sirois, 2009; Sirois & Jackson, 2012; Zhang et al. 2018). These studies all suggest a tight relationship between pupil change and prediction error, such that the larger the prediction error, the larger the pupil change. Furthermore, not only is this relationship between pupil change and prediction error supported by behavioral evidence, it is also linked to the PH Model (Koenig et al., 2018) and the RW Model (Zhang et al., 2018, see above for more descriptions of these prediction error-based learning models). In a series of three experiments by Koenig et al. (2018), participants learned to associate different cues with

different outcomes. All experiments were designed to examine pupil change during processing of cues that elicited no prediction error (i.e. cues that were consistently followed by the outcome) versus cues that elicited prediction error (i.e. cues that were sometimes not followed by the outcome). Results revealed a close match between pupil change and the predictions of the Pearce-Hall model: During early learning trials, no prediction error cues caused larger pupil change than prediction error cues, but this effect reversed such that towards the end of the experiment, the prediction error cues caused a larger pupil dilation.

Zhang et al. (2018) designed a study where they used pupillometry data in conjunction with the RW model to investigate prediction error in both infants and adults. All participants completed an identical implicit learning task designed to establish associations between two pairs of sounds and images. After a period of familiarization, participants were introduced to error trials (i.e. the sound was not followed by the paired image). By taking a computational approach, the trial-by-trial model output revealed changes in a participant's expectation and prediction error. Specifically, they found that infant and adult's pupil change to the images significantly correlated with the model's estimation of prediction error. Importantly, they also found convergence of the participants' expectation and prediction error to the structure of the task. These results suggest that pupil size closely tracks the prediction error established in the RW model.

These studies often examine the subcomponents of prediction in isolation, however, we want to investigate more than one subcomponent at the same time, and specifically to look at the interrelationship between prediction error and anticipation. To do this, we expanded the paradigm used by Reuter, Emberson, Romberg & Lew-Williams (2018), a paradigm that measures infants' ability to anticipate and shows updating of anticipation. While this paradigm

was excellent for using AEMs to investigate infants' attention, it doesn't allow the measure prediction error, despite having a design that should elicit a prediction error. In this study, Reuter et al. (2018) tested 12- to 24-month-old infants and related anticipation in the visual domain to an infant's progress in language development. Specifically, they investigated the relationship between a visual prediction task designed to elicit AEMs as a measure of anticipation and had parents fill out the MCDI to measure infants' vocabulary size. The visual prediction task was composed of two blocks with 8 identical trials per block. In Block 1, the target appeared on one side of the screen and in Block 2, the target appeared on the opposite side. After controlling for differences in infants' speed of visual processing, Reuter et al. (2018) found no correlation between proportion of AEMs and vocabulary size. Instead, they found infants with larger vocabulary size updated anticipations more readily. Specifically, by analyzing infants' AEMs in Block 2, they found a significant correlation between proportion of AEMs to the novel location and vocabulary size. Furthermore, Reuter et al. (2018) divided infants into high-vocabulary and low-vocabulary groups based on a median split in MCDI scores and examined the trial-by-trial change in proportion of AEMs to the novel location in Block 2. Indeed, the authors found that the high vocabulary group made more AEMs to the novel location compared to low vocabulary group on a trial-by-trial basis. Taken together, Reuter et al. (2018) determined that it is the ability to update anticipations based on new information, over the ability to generate more anticipations overall, to be important for vocabulary growth. In other words, the ability to update anticipations and the experience of prediction error may be linked in this study and have important consequences for learning, however, Reuter et al. (2018) was unable to answer this question because they did not directly measure prediction error (for studies that integrate across

the subcomponents of prediction, specifically prediction error and updating, see Kayhan, Hunnius, O'Reilly & Bekkering, 2019; Kayhan, Meyer, O'Reilly, Hunnius & Bekkering, 2019).

The current study aims to build upon Reuter et al. (2018) to investigate how prediction error influences the ability to update anticipations (i.e. learn) by incorporating a direct measure of prediction error, pupillometry, into the task, without altering the task. We hypothesized that if prediction error is the spark needed to update our anticipation based on new information, then infants with larger pupil change, or greater experience of prediction error, will generate more correct AEMs, or update more readily. Overall, to investigate the relationship between the experience of prediction error and the ability to update anticipations, we have expanded the approach of Reuter et al. (2018) to include measures of both AEM, as our measure of anticipation, and pupillometry, as our measure of prediction error.

Experiment 6: Infants

In Experiment 6, we wanted to identify whether prediction error is the mechanism by which infants know to update anticipations. We used pupil size to measure prediction error and correct AEMs to measure infants' ability to update anticipations. Specifically, we examined how pupil size during trials when the target switches to a novel location, which should generate prediction error, relate to the proportion of correct AEMs, which indicates infants' ability to update, for the same trials. To this end, we tested 14-month-old infants in an extended version of the same visual prediction task as Reuter et al. (2018).

Methods

Participants

We collected data from 41 infants. Three infants were excluded due to insufficient number of trials as a result of movement; four infants were excluded for not paying attention (i.e. looking at the screen for less than half of the experiment), two infants were excluded due to technical errors, and one infant was excluded for being outside the age range of 14 months old. With these infants excluded, our final infant sample included 31 infants (M = 14 months 17 days; SD = 9 days; sixteen females). Infants in the sample were full-term at birth, defined as greater than or equal to 37 weeks gestation, with no reported complications or medical/developmental problems. Experiments were approved by the university's Institutional Review Board and informed consent was obtained before beginning the study from the legal guardian of the infant. Families received $10, a t-shirt, and a children's book for their participation.

Stimuli

The experiment was made up of 3 blocks. Each block was composed of identical center fixation (a looming pinwheel) but a different pairing of target stimulus (circle, square or triangle) and sound (xylophone, laser sound, magic sound) were used for each block. All visual stimuli and background color of the screen were made isoluminant to each other and the room to prevent changes in pupil size as a result of varying brightness. We used an isoluminance script provided by Daniel Swingley (personal communication, David Brainard and Daniel Swingley, 2016) which takes a standard RGB image, corrects the gamma, separates the image's chromaticity and luminosity, and sets the luminance to a constant value, thereby making the image isoluminant, without affecting the chromaticity. We selected a luminance value of 0.3 to match the lighting in the room and not distort the visual stimuli and videos to the point of being unrecognizable.

Procedure

Infants sat on their caregivers' laps throughout the experiment. Caregivers were instructed to not interfere with the infant throughout the study and wore a sleeping mask during the experiment, which prevented them from seeing the screen and blinding them to the content of the individual trials (i.e., to prevent biasing of infant behaviour). Adult participants sat approximately 60 cm from the display monitor measured 34cm by 27cm and eye gaze was recorded at 250Hz (i.e. one data point collected every 4 ms) using the Eyelink 1000 eye tracker (Eyelink 1000, SR Research, Ontario, Canada) and the 25 mm infant lens located below the monitor. Stimuli were presented using MATLAB for Mac (R2015b) and Psychtoolbox (3.0.12 Beta, SVN revision 6399).

The host computer, which runs the Eyelink software, plus experimenting computer, which runs MATLAB, were in front of the experimenter. The experimenter watched the participant from the host monitor in order to exit the experiment when infants became too inattentive and fussy. Monitoring the host computer also allowed the experimenter to adjust the display monitor as infants moved. A three-point calibration was used because we were interested in where participants looked along the horizontal coordinates (e.g. left, right, center). We performed calibration and validation for all participants and only included participants that had good (i.e. the highest level) validation.

The current paradigm is an extension of Reuter et al., (2018). We had participants complete the same task but three times, for a total of 3 blocks. We extended the experiment because pupil change is a noisy measurement, especially for infants, and we wanted to average across the three blocks for a stable prediction error measure. Each block was made up of 8 identical pre-switch trials (trials 1 – 8) followed by 8 identical post-switch trials (trials 9 – 16),

for a total of 16 trials per block. Every trial began with a blank period of 250 ms to allow for infants to disengage from previous target, then center fixation for 1500 ms, followed by a delay for 1200 ms, and finally the target for 1000 ms (**Figure 30).** The delay for the current experiment is longer than Reuter et al. (2018) because our infants were, on average, younger, and we wanted to give them enough time to elicit an AEM. Whether the target showed up on the left or right side of the screen was randomized for pre-switch, and the target would show up on the opposite side for post-switch.

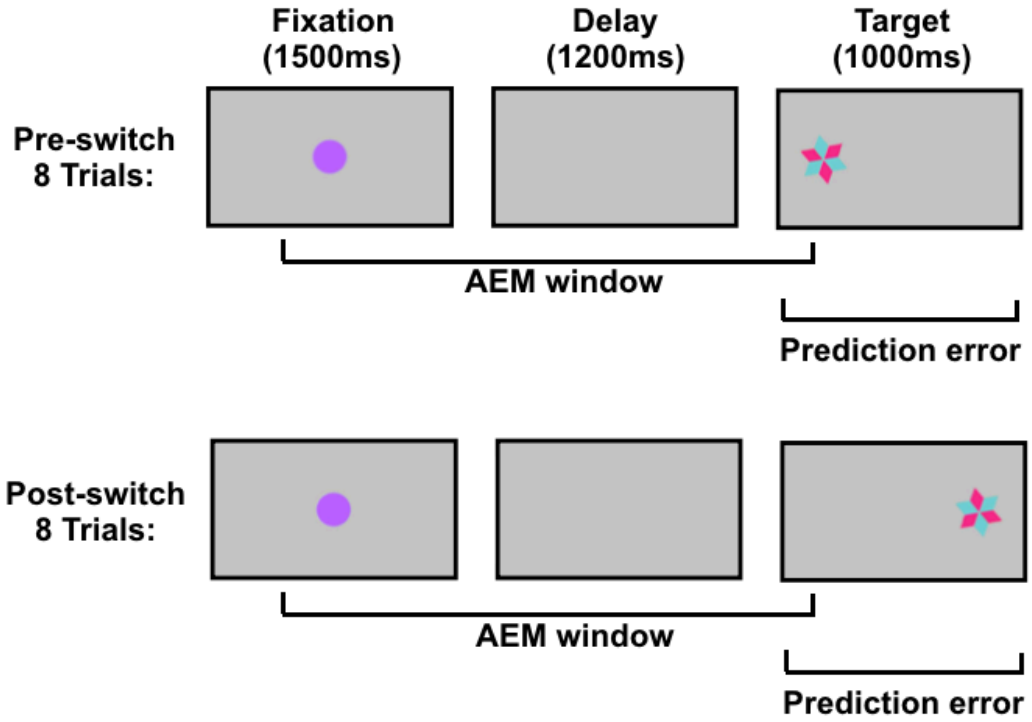


Figure 30. Schematic of one block in the experiment. Participants saw a total of three blocks in the experiment. For each block, participants saw 8 identical pre-switch trials where the target appeared on one side (e.g. left), followed by 8 identical post-switch trials where the target appeared on the opposite, novel side (e.g. right). For each trial, we measured anticipatory eye movements (AEM) during the window of 400 ms after center fixation onset to 200 ms after target onset and prediction error as average pupil change during target period (2950ms – 3950ms).

Preprocessing

Data Preprocessing

Data files exported from the eye tracker were processed to determine pupil change for each participant. Blinks were identified and linearly interpolated (Hepach & Westermann, 2013; Addyman, Rocha & Mareschal, 2014; Fawcett, Wesevich & Gredebäck, 2016) using the package zoo (Version 1.7-13, Zeileis, Grothendieck, Ryan & Andrews) for the R software (Version 3.2.1; R Development Core Team, 2015). Samples collected when participants were looking away from the screen (i.e. X and Y coordinates were outside the bounds of the screen dimension, 1280 x 1024) were removed.

Eye Movements

Each sample of eye movement data collected was coded as looking at upcoming/current target location, center of the screen, previous target location or missing based on the X coordinate. Eye movements were considered to be anticipatory if participants first looked at center fixation, and initiated eye movement to the upcoming/current target location between 650 ms – 3150 ms of a trial (from 400 ms after center fixation onset to 200 ms after target onset). Our AEM window is larger than Reuter et al. (2018) who used a more conservative time window of 200 ms before center fixation offset until 200 ms after target onset. We expanded our AEM window by including more time during the period after center fixation but before target onset in order to give infants more time to make an AEM because they were younger.

Once a look is coded as an AEM (i.e., landing on a target location in an anticipatory fashion, see above for more details), we coded each AEM as correct or incorrect based on the location of their eye movement. A correct AEM is an AEM that lands on the upcoming target location and an incorrect AEM is an AEM that lands on the opposite location of the target. All

analyses only included the first anticipatory shift, following Romberg and Saffran (2013) and Reuter et al. (2018).

Pupil Size

Pupil size was our measure of prediction error on each trial. We analyzed pupil size by calculating percentage change from baseline, which was the first 1950 ms of each trial. We chose this period as baseline because participants viewed a center fixation and it was identical for every trial. The baseline measure was subtracted from each pupil diameter measurement, divided by baseline, and multiplied by one hundred, to get percentage change from baseline. We used the average pupil change during the target period (2950 ms – 3950 ms) of post-switch trials as our measure of prediction error because this is the period where participants experience the strongest prediction error. We also excluded trials with average pupil change larger than 2.5 standard deviations from the mean during the stimulus period for each subject, which is likely a result of eye movement.

Results

Anticipation (AEMs)

We started by confirming that infants were performing the task properly by determining whether or not they were generating anticipations (i.e. AEMs). We excluded trial 1 of each block from all analyses on the assumption that infants are not or cannot be generating anticipations on this trial. Using the remaining trials, we conducted a 2 x 2 repeated measures ANOVA on proportion of anticipations, with AEM type (correct, incorrect) and switch (pre-switch, post-switch) as the within, repeating factors (**Figure 31**). This analysis revealed a significant interaction $F(1, 103) = 5.78$, $p = .018$, but no main effect of AEM type ($F(1, 103) = 0.18$, p

= .67) or block ($F(1, 103) = 1.04, p = .31$), suggesting infants generated more correct AEMs on pre-switch trials and more incorrect AEMs on post-switch trials.

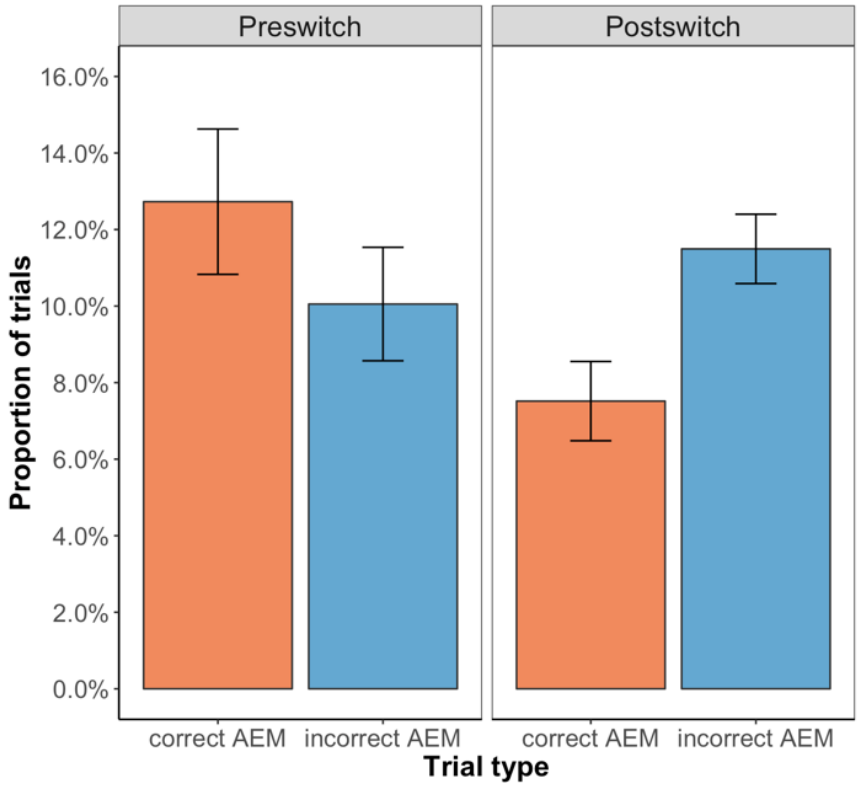


Figure 31. Proportion of trials that had correct or incorrect AEMs, separated by pre-switch and post-switch for infants. Vertical error bars indicate between block SEM.

In addition, we conducted an analysis of the post-switch trials to confirm that infants experienced prediction error in the early post-switch trials, but learned and updated their anticipations later on (**Figure 32**). We performed a linear regression using trial number in post-switch and trial type (correct AEM or incorrect AEM) as the independent variables and proportion of infants as the dependent variable. We found a significant interaction between trial number in post-switch and trial type such that with each exposure to a post-switch trial, the proportion of infants that made an incorrect AEM decreased by 3.57% ($B = -0.0357, p = 0.0011$). Analysis of how the proportion of correct and incorrect AEMs changed trial by trial for pre-

switch trials are in Appendix I. Overall, these analyses suggest infants are responding to the structure of the task, especially during post-switch where more infants made incorrect AEMs during the early trials but updated their anticipations and made more correct AEMs during the later trials.

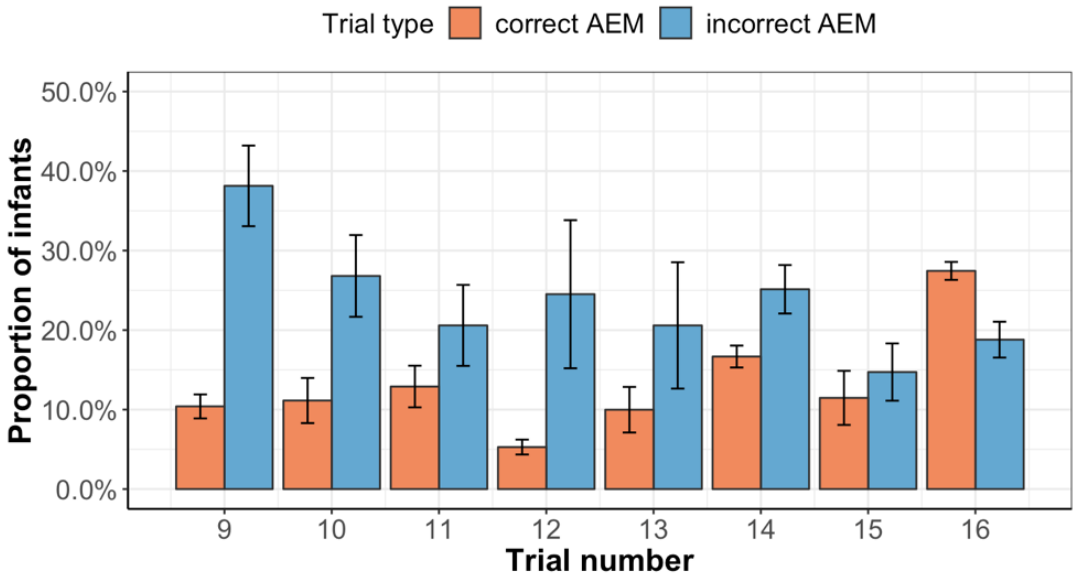


Figure 32. Proportion of infants that generated a correct AEM versus incorrect AEM on a trial by trial basis during post-switch when the target appeared in the novel location. Vertical error bars indicate between block SEM.

Prediction Error (Pupil Change)

Before discussing the pupil results, it is important to address an issue that often occurs when eye movements and pupil change are measured in the same study. Namely, studies have shown that pupil size can be distorted as a result of eye movement. This phenomenon is well documented by Hayes and Petrov (2015) when they measured the size of an artifical eye, which has a fixed pupil size, and found pupil size recordings increasing when the artifical eye scanned the screen from left to right. We confirmed the presence of this exact pupil distortion within data collected from the current study (Appendix J). Although researchers have put forth approaches

for correcting this pupil distortion, these correction techniques are not infant friendly because they require numerous pre-experiment measurements and rely on participants staying completely still throughout the experiment (Gagl, Hawelka, Hutzler, 2011; Hayes & Petrov, 2015). Nevertheless, studies have been successful in integrating eye movements and pupil size despite these methodological challenges and without these correction techniques (Cheng, Kaldy, Blaser 2019; Kleberg, del Bianco, Falck-Ytter, 2019; Ross-Sheehy & Eschman 2019).

Focusing on the current study, we have corrected for these potential biases in pupil size in a number of convergent ways: First, we carefully selected a baseline: The first 1950 ms of each trial where participants viewed a center fixation. Crucially, this aspect of the experiment was identical for every trial, had no or very minimal eye movements and a central fixation. This period allows for stable and non-distorted pupil measurements. Then, pupil change was calculated from baseline rather than relying on raw pupil size. Second, we counterbalanced the location of the stimuli that appeared first in each block. This is important to ensure that any distortions of pupil size dependent on position of the stimulus is not confounded with our experimental manipulations. Since pupil size measurements become larger when looking at the right side of the screen compared to the left, counterbalancing the location of the stimuli is necessary to ensure that there is no systematic relationship between stimulus location and our experimental manipulations like stimuli post-switch that are meant to elicit prediction error. Overall, we have implemented a number of steps to minimize or eliminate any distorting effects of eye movements on pupil size.

Having establishing that 14-month-old infants are exhibiting eye movements appropriate to the task (above), we now consider pupil change, our measure of prediction error. As mentioned previously, our data indicates that infants make more incorrect AEMs during post-

switch compared to pre-switch. Thus, if pupil change is a valid measure of prediction error in this task, we hypothesize to find larger pupil change for post-switch compared to pre-switch. We calculated average pupil change during the target period, and did not find a significant difference in pupil change between pre-switch and post-switch trials (within subjects, $t(30) = 1.12$, $p = 0.27$, **Figure 33**). We also plotted the time-course of pupil change for pre-switch and post-switch trials and conducted an exploratory analysis to determine whether the time window we chose was incorrect (**Figure 34**). We conducted two-tailed t-tests across successive time points and corrected for multiple comparison using a conservative method (the Benjamini & Hochberg, 1995; procedure controlled for the rate of false discovery; Mill, O'Connor & Dobbins, 2016; Zhang et al., 2018; and found no period of time that was significantly different pupil responses across trial types ($p > 0.05$). In other words, throughout the entire trial, there was no time period where the pupil change was larger for post-switch compared to pre-switch trials.

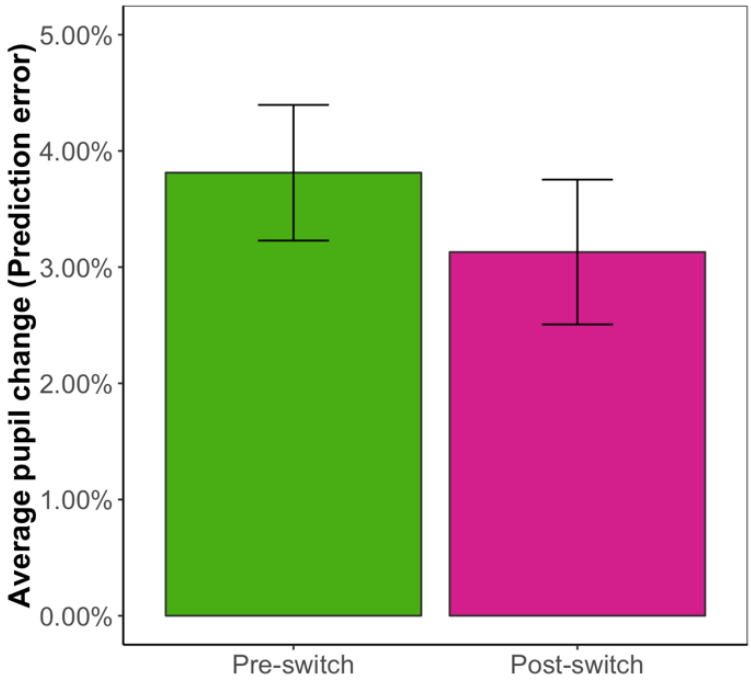


Figure 33. Average pupil change during target period for pre-switch versus post-switch trials for infants. Error bars represent between subject SEM.

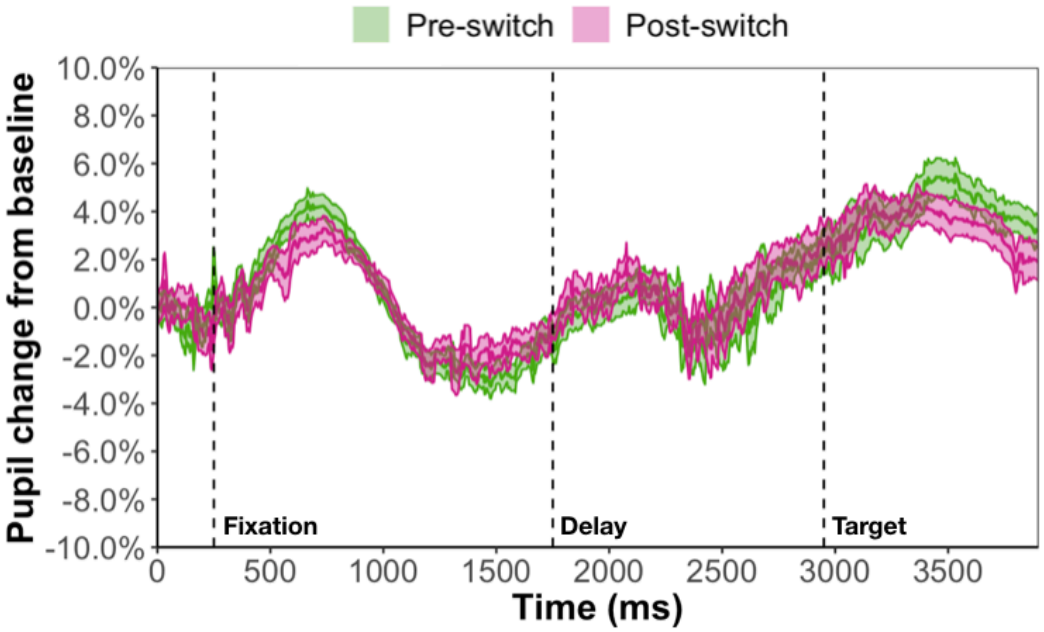


Figure 34. Time-course of pupil change for pre-switch trials and post-switch trials for infants.

We conducted further investigation of our prediction error measure. We compared average pupil change for the last trial of pre-switch, when the target should generate minimal prediction error, versus the first trial of post-switch, when the target appears in the novel location and, we reasoned, prediction error would be the largest. This analysis, however, revealed no significant results ($t(28) = 1.22$, $p = 0.23$, **Figure 35**).

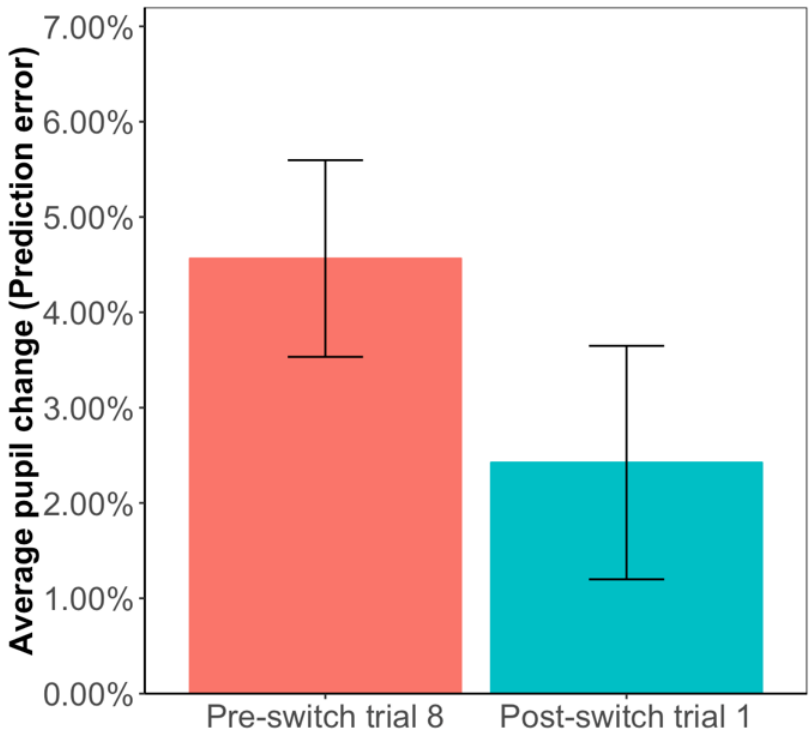


Figure 35. Average pupil change during target period for pre-switch trial 8 versus post-switch trial 1 for infants. Error bars represent between subject SEM.

We also explored whether this effect is being reduced across blocks. A major difference between the current experiment and Reuter et al. (2018) is their experiment was composed of 1 set of pre-switch and post-switch trials whereas our task was composed of 3 sets. Comparing average pupil change between pre-switch trial 8 and post-switch trial 1 separately for each block also revealed no significant interaction ($F(2, 113) = 1.25, p = 0.29$), no main effect of block ($F(2, 113) = 0.091, p = 0.91$) but a significant main effect of trial ($F(1, 113) = 5.15, p = 0.025$). Furthermore, results were in the wrong direction with greater pupil responses to pre-switch compared to post-switch trials (**Figure 36**). Given the opposite direction of results we found, we decided to replicate this study with adults, because we have more confidence in the quality of adult data collected and there is a larger body of research suggesting a link between prediction

error and larger pupil change for adults (Koenig et al., 2018; Nassar et al., 2012; O'Reilly et al., 2013; Zhang et al., 2018).

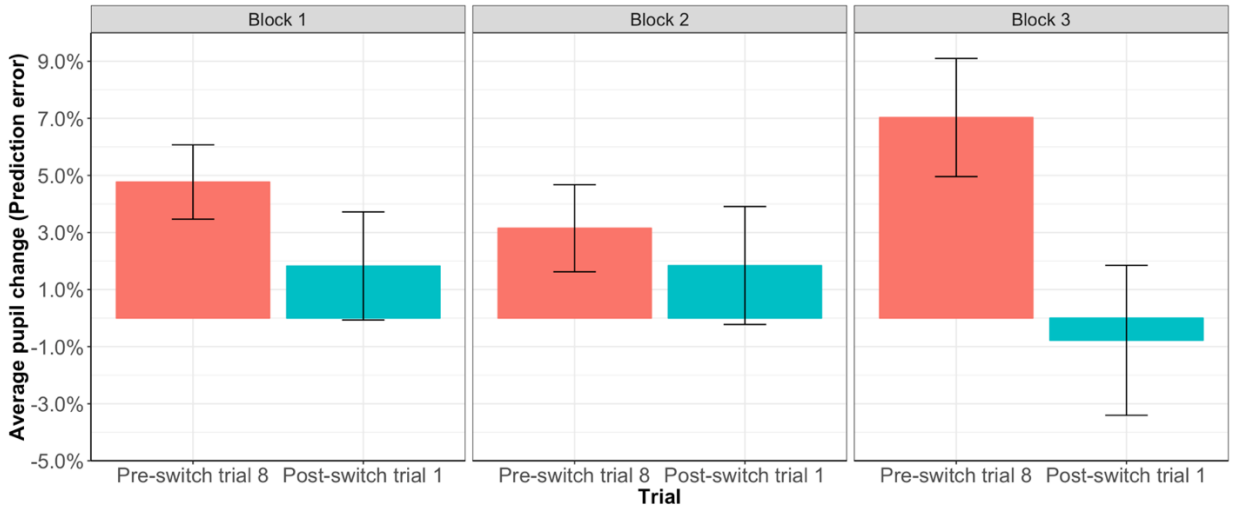


Figure 36. Average pupil change during target period for pre-switch trial 8 versus post-switch trial 1 for each block for infants. Error bars represent between subject SEM.

Discussion

Overall, we found reliable measures of AEMs in this study, suggesting infants are learning the spatial pattern, enabling them to predict upcoming stimulus. However, we did not find significant pupil change results, which prevented us from conducting further analyses of prediction error, and the relationship between prediction error and updating. Thus, in Experiment 7, we conducted an identical study with adults to examine the lack of pupil change results.

Experiment 7: Adults

In Experiment 7, we examined the relationship between prediction error (average pupil change) and updating (proportion of correct AEMs), in healthy adults, using the same procedure as Experiment 6. We were especially interested in the pupil change results because we want to determine whether or not the lack of pupil change results from Experiment 6 was a result of the

population we were testing (i.e. infants) or a larger methodological issue. If the lack of results is infant specific, we expect to find prediction error related pupil differences in adults, but if the lack of results is a suggestion of a broader issue, then it is possible we will also find no pupil difference in adults.

Methods

We collected data from 31 adults (M = 19.6 years; SD = 2.3 years; twenty-four females) in order to match the number of infants. One additional adult was excluded for being outside of the age range and two additional adults were excluded due to technical errors. Experiments were approved by the university's Institutional Review Board and informed consent was obtained before beginning the study from the adult participant. Adult participants received $8 for their participation.

Results

Anticipation (AEMs)

We first confirmed that adults were paying attention to the study by determining whether or not they were generating anticipations (i.e. AEMs), as this was a study designed for infants and might not be interesting for adults. We excluded trial 1 of each block from all analyses. We conducted a 2 x 2 repeated measures of variance (ANOVA) on proportion of anticipations, with AEM type (correct, incorrect) and switch (pre-switch, post-switch) as the within, repeating factors (**Figure 37**). This analysis revealed a main effect of AEM type $F(1, 100) = 55.93$, $p < 0.001$, no main effect of block ($F(1, 100) = 0.52$, $p = 0.47$) or significant interaction ($F(1, 100) = 0.043$, $p = 0.84$) suggesting adults make more correct AEM than incorrect AEMs regardless of

pre-switch or post-switch, which is possibly a result of how quickly adults can adjust their anticipations in this simple task.

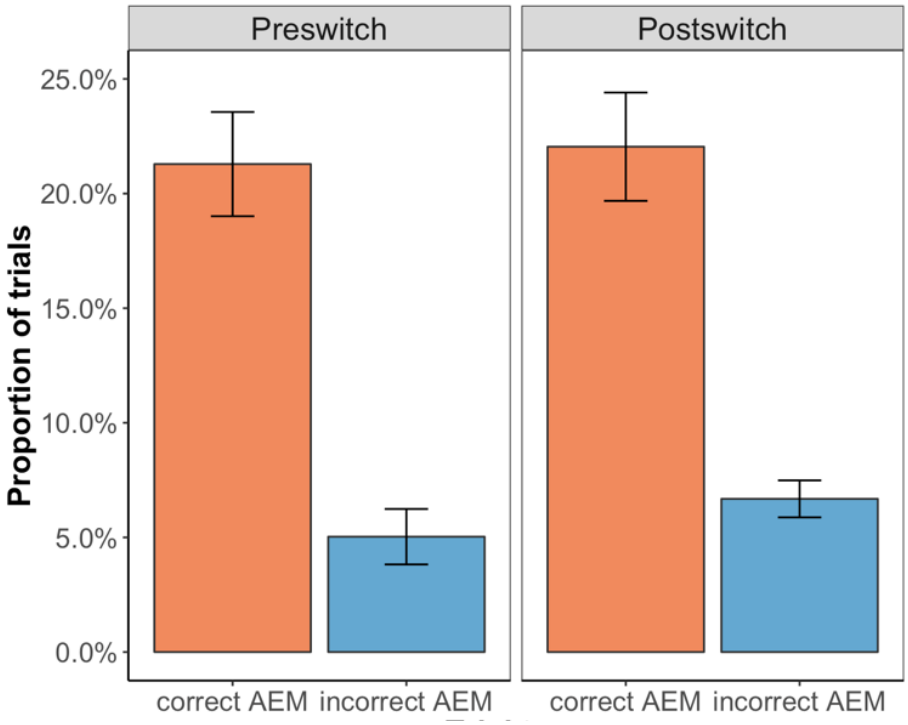


Figure 37. Proportion of trials that had correct or incorrect AEMs, separated by pre-switch and post-switch for adults. Vertical error bars indicate between block SEM.

Next, we examined how the proportion of correct and incorrect AEM changes on a trial-by-trial basis for post-switch trials (**Figure 38**). We performed a linear regression using trial number in post-switch and trial type as the independent variables and proportion of adults as the dependent variable. We found a significant interaction between trial number in post-switch and trial type such that with each exposure to a post-switch trial, the proportion of adults that made an incorrect AEM decreased by 7.7% (B = -0.0743, $p < 0.001$; **Figure 38**). Analysis of how the proportion of correct and incorrect AEMs changed trial by trial for pre-switch trials are in Appendix K.

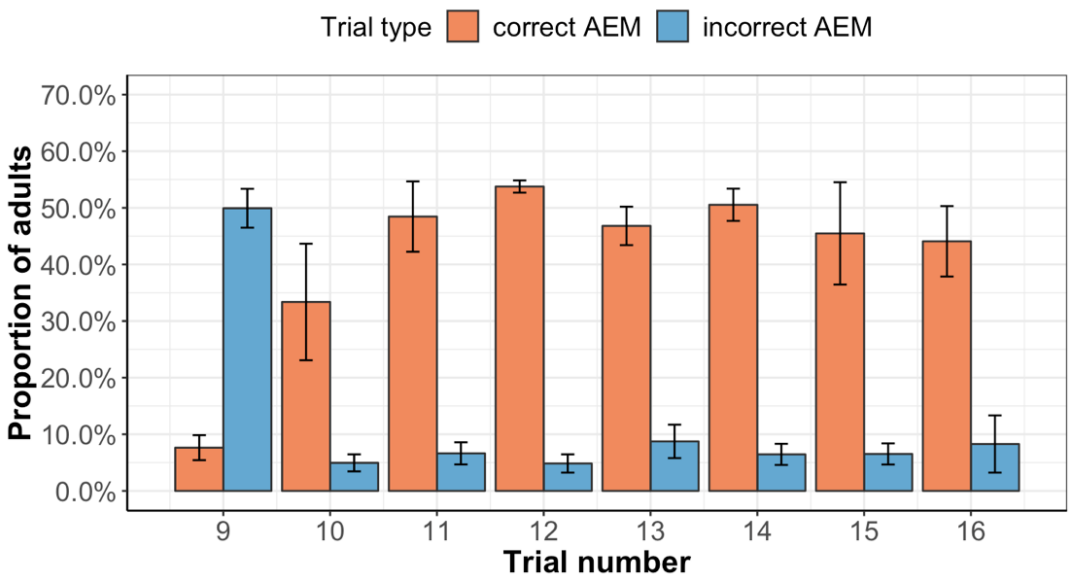


Figure 38. Proportion of adults that generated a correct AEM versus incorrect AEM on a trial by trial basis during post-switch when the target appeared in the novel location. Vertical error bars indicate between block SEM.

Prediction Error (Pupil Change)

After establishing that adults are exhibiting AEMs appropriate to the task, albeit at a much faster rate, we moved onto pupil change. First, we confirmed the presence of pupil distortion within the adult data (Appendix L). Next, unlike Experiment 6, we did not compare average pupil change for pre-switch and post-switch because adults did not generate more incorrect AEMs in post-switch. As per our previous analysis, we found that adults make similar proportion of correct and incorrect AEM for both pre-switch and post-switch. Therefore, we wouldn't expect pupil change to be different between pre-switch and post-switch. Instead, we compared average pupil change for the last trial of pre-switch versus the first trial of post-switch, which we know from **Figure 38**, generates the highest number of incorrect AEMs. This analysis also revealed no significant results ($t(30) = 0.29$, $p = 0.78$, **Figure 39**). We also examined the timecourse of pupil change for pre-switch trial 8 and post-switch trial 1 to determine if the

prediction error response appeared in a different time period (**Figure 40**). We conducted two-tailed t-tests across successive time points and corrected for multiple comparison using the Benjamini & Hochberg procedure (Benjamini & Hochberg, 1995) and found no period of time had significantly different pupil responses between pre-switch trial 8 and post-switch trial 1 ($p > 0.05$).

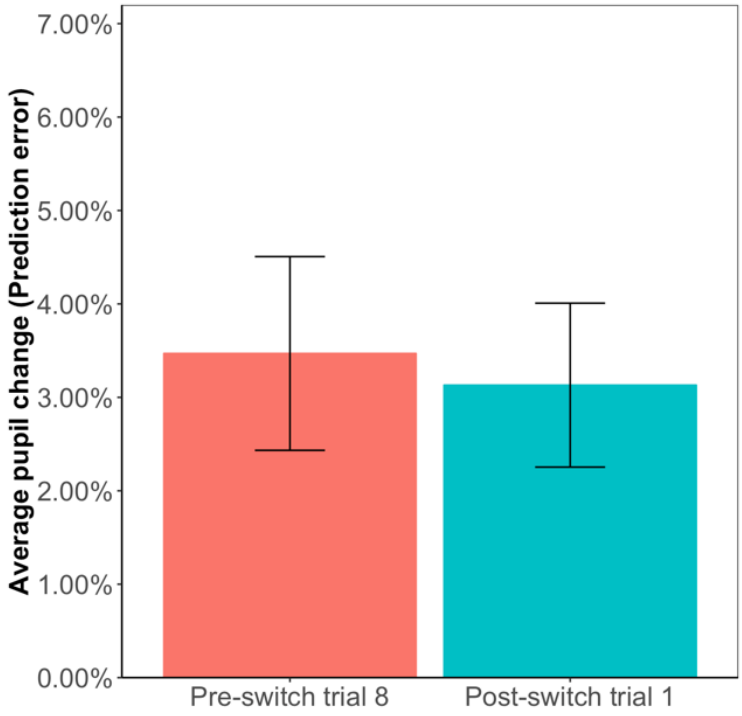


Figure 39. Average pupil change during target period for pre-switch trial 8 versus post-switch trial 1 for adults. Error bars represent between subject SEM.

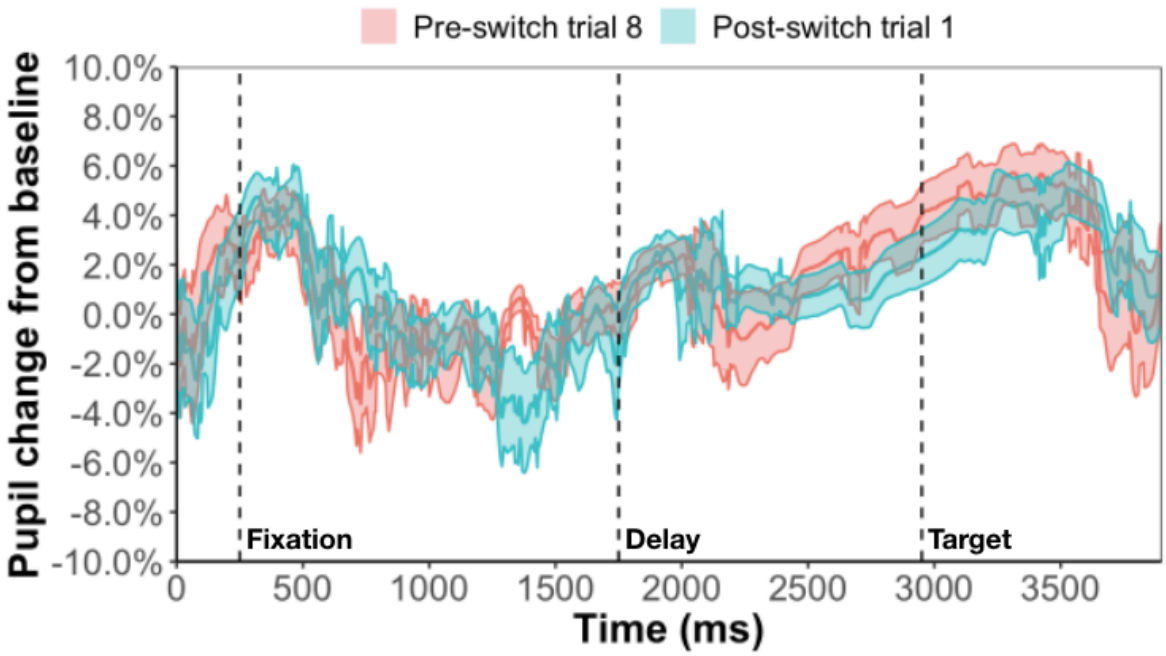


Figure 40. Timecourse of pupil change for pre-switch trial 8 and post-switch trial 1 for adults.

The lack of pupil change results were unexpected given the numerous studies suggesting a connection between larger pupil change and prediction error in adults (Koenig et al., 2018; Nassar et al., 2012; O'Reilly et al., 2013; Zhang et al., 2018). Thus, we conducted an exploratory analysis to examine the average pupil change for pre-switch trial 8 and post-switch trial 1 for each block (**Figure 41**). We performed a 2 x 3 ANOVA on proportion of anticipations, with trial (pre-switch trial 8, post-switch trial 1) and Block (1, 2, 3) as the within, repeating factors, which revealed no significant interaction ($F(2, 178) = 0.47, p = 0.63$) or main effects (trial: $F(1, 178) = 0.008, p = 0.93$; block: $F(2, 178) = 0.17, p = 0.85$). This block by block analysis suggested that pupil change did not reflect prediction error in any block.

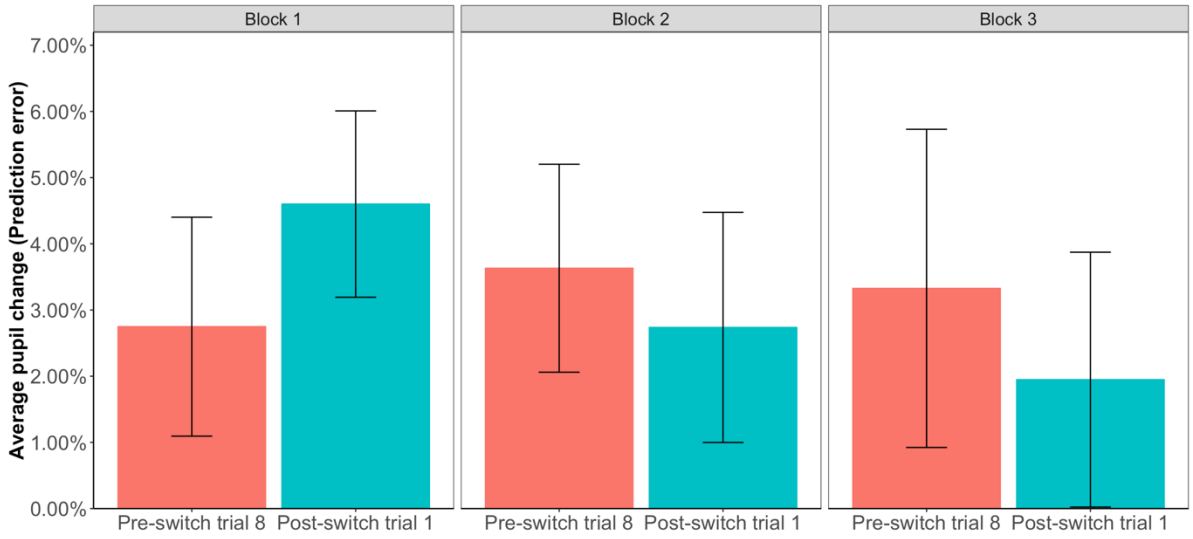


Figure 41. Average pupil change during target period for pre-switch trial 8 versus post-switch trial 1 for each block for adults. Error bars represent between subject SEM.

Discussion

In Experiment 7, we replicated Experiment 6 with adults, to determine whether the lack of difference between pre-switch and post-switch pupil change, was a result of testing infants, who tend to have noisier data, or a result of a larger issue with pupillometry. First, we determined that adults were generating eye movements that suggest they were performing the task properly. However, when we analyzed pupil change and compared pre-switch trial 8 with post-switch trial 1, where we expected to find the largest pupil difference, we did not find a significant difference. It is interesting to note that when we examined the average pupil change for pre-switch trial 8 and post-switch trial 1 for each block, the direction of the pupil change is correct for block 1 but not block 2 and block 3. Given the simplicity of the task and the speed at which adults learned, it is possible that the switch of the target to the opposite location, did not induce a strong prediction error response in adults. As a result, we saw pupil change in the right direction for block 1, but quickly disappeared for block 2 and block 3. Overall, we believe the lack of significant results linking pupil change to prediction error for adults in this task, is a

result of the prediction error being too weak for adults. We discuss what this might entail for future research in the overall discussion.

General Discussion

Prediction plays an important role in infant and adult learning. Although prediction is often studied as a single process, it is composed of several subcomponents: anticipation, expectation, prediction error and updating (Zhang, 2020). Research investigating prediction has moved from considering prediction as a single process to investigating these subcomponents (de Lange et al., 2018; Summerfield & Egner, 2009; den Ouden et al., 2012) However, previous work has predominantly investigated them in isolation, resulting in a fragmented understanding of prediction rather considering these processes as an integrated whole. Thus, the current studies were designed to investigate two crucial subcomponents of prediction (anticipation, prediction error) using two different dependent measures with the goal of interrelating the two to determine how infants use prediction errors to update anticipations.

In Experiment 6, using the same paradigm as Reuter et al. (2018), 14-month-old infants completed a visual prediction task as we measured their eye movements and pupil size. We found that infants generated AEMs, suggesting they learned the spatial pattern of the task, and used this pattern to generate anticipations of future stimuli. Furthermore, during post-switch trials (i.e. after the target switches to the opposite side), infants also gradually updated their anticipations. In this way, we replicated the main finding from Reuter et al. (2018). In Experiment 7, we expanded this paradigm to an adult population. When we conducted the same analysis for adults, we found adults generated the correct anticipations and updated their anticipations, as with the infants, but did so much faster than infants. This is to be expected as

adults were completing a task designed for infants, which allowed them to learn the spatial pattern of the task and adjust to the switch at a much faster rate. Overall, the eye movement results found across the two studies suggest infants and adults are capable of generating and updating anticipations.

When we analyzed the pupil change (the index of prediction error), however, there was no evidence for prediction error or larger pupil change when the target appeared at the opposite location for both infants (Experiment 6) and adults (Experiment 7). Overall, we fail to find that this paradigm produces a good measure of prediction error. As a result, we conducted a number of exploratory analyses which revealed interesting similarities and differences across the two studies that are worth discussing.

First, we conducted an exploratory analysis to determine whether we selected the correct time window to analyze pupil change for both infants and adults. This analysis revealed that at no point during the trial was pupil size larger for prediction error. Previous studies find significant pupil difference related to prediction error starting at approximately 1000ms after target presentation (Braem, Coenen, Bombeke, van Bochove, & Notebaert, 2015; Chen & Westermann, 2018; Wetzel, Buttelmann, Schieler, & Widmann, 2016; Zhang et al., 2018), whereas our trial ends exactly 1000ms after target presentation. Thus, it is possible that if we extended our target period we might see a pupil difference. However, we believe that even with an extension of target period we will not see a pupil difference because the previous studies show dissociation of the conditions in the right direction much earlier than 1000ms (Braem et al., 2015; Chen & Westermann, 2018; Wetzel et al., 2016; Zhang et al., 2018), and we do not see the hypothesized dissociation of conditions in our experiments even in this earlier time window. Also, previous adult studies find that the adult pupil response diverges sooner than infants (e.g.,

Zhang et al., 2018 for a direct comparison), and we do not observe any divergent in our adult sample. Taken together, these results indicate that the lack of pupil change difference is not a result of time window selection or a short window for analysis, suggesting perhaps a separate issue with the measures and task that were used, which we will discuss in more detail below.

Second, when we conducted a block analysis of average pupil change for pre-switch trial 8 and post-switch trial 1, which is where we expected to see the largest pupil difference because pre-switch trial 8 should generate the least amount of surprise whereas post-switch trial 1 should generate the most amount of surprise, we found that the infant data revealed larger pupil change for pre-switch trial 8 for almost all the blocks. When we conducted the same analysis for adults, however, we found an interesting pattern of results suggesting adults may have experienced prediction error in block 1, but that effect disappeared in block 2 and block 3. We believe the effect for adults disappeared because the experimental manipulation used in this task was too transparent or easy for adults to be surprised each block. In other words, the adults likely learned the metapattern of the experiment and, thus, were not surprised when the target changed location in blocks 2 and 3. This line of thinking is further supported by the evidence showing how quickly adults learn in this task. Comparing infants and adults in this exploratory analysis, we suspect that finding positive evidence for prediction error in the first block for adults and not for infants is not likely to be a result of power issue because the same number of infants and adults contributed data for block 1. Thus, the remaining possibilities are poor data quality from the infants resulting in noisier signal and/or pupil change reflecting a different cognitive process other than prediction error in this task, which is discussed below. Overall, across the two studies, despite eye movements suggesting participants were generating anticipations and learning to update, and the methodological approaches we implemented to eliminate pupil distortions, pupil

change did not reflect prediction error in this task, which prevented us from analyzing the relationship between prediction error and updating.

Why did pupil change not reflect prediction error in this task? Pupil change has been used to measure a variety of cognitive processes other than prediction error (*cognitive control:* van der Wel, & van Steenbergen, 2018; *attention:* Unsworth & Robison, 2017; *emotion:* Geangu, Hauf, Bhardwaj, & Bentz, 2011, *memory:* Goldinger & Papesh, 2012). One interesting line of work are studies demonstrating larger pupil change as a sign of familiarity (Kafkas & Montaldi, 2012, 2015) which is the opposite of surprise. With respect to the current study, we expected pupil change to decrease with each additional pre-switch trial as they become less surprised, but it is possible that in this paradigm, pupil change actually increased because they are focusing on the familiarity of the stimuli, resulting in a larger pupil change on the last trial of pre-switch. Subsequently, when the first post-switch trial occurs, given the sudden change of location, pupil change becomes more reflective of surprise than familiarity, which also results in a larger pupil change. Thus, when we computed pupil change between the last pre-switch trial and the first post-switch trial as a group average, and compared pupil change average for pre-switch and post-switch trials in infants, they are comparable.

It is important to note that pupil change did not reflect familiarity in Zhang et al. (2018), and we hypothesize it is a result of the differences in task design. In Zhang et al. (2018), each trial was composed of an auditory cue and blank screen, followed by a visual stimulus in the center of the screen. Compare that trial set up to the current study, where each trial is composed of a center fixation paired with a specific sound, followed by a blank screen, then the target stimulus appeared on the left or right side of the screen. Given the simplicity of each trial and overall experiment design in Zhang et al. (2018), it is possible that infants adjusted to the task

quickly, which is supported by the trial by trial changes in prediction error and prediction modelling results. Behaviorally, this quick adjustment is reflected by a smaller pupil change, and consequently, a larger pupil change when the prediction error occurred.

In summary, we replicated the anticipatory eye movements found in Reuter et al. (2018) and expanded this paradigm to also consider pupil change. Comparing adult and infant samples, we found interesting similarities and differences between infants and adults in their AEMs and pupil change. However, we do not find the hypothesized direction of pupillometry results in either group, and we believe that pupil change in this task was perhaps capturing something other than prediction error. These findings occurred despite including a number of approaches to eliminate the bias in pupillometry associated with looking to different areas of the screen. Nevertheless, given the theories and computational models supporting the importance of prediction error for updating, we believe future studies should continue to pursue this area of research.

Conclusion

Using a combination of pupillometry, eye movements, and computational learning models, the chapters presented in this dissertation contribute to our current understanding of prediction and pupillometry as a method for developmental research in several ways. First, our emphasis on the breakdown of prediction into its component parts, namely anticipation, expectation, prediction error and updating, is important for the goal of applying prediction, as a mechanism, to infant learning and development (Chapter I). Second, we also demonstrated the importance of examining multiple subcomponents of prediction in one study because this approach is the only solution for understanding how the subcomponents work together in supporting predictive processing in infancy as a whole (Chapters IV and Chapter V). Third, we have provided strong evidence that pupillometry can be used to expand our understanding of cognitive development given its high temporal resolution, compatibility with participants across the life span, and ability to work with other behavioral measures and computational models. Specifically, we have shown that pupil size is a good marker of prediction error and fitting pupil size to a classic error-driven learning model revealed that 6-month-old infants and adults have similar predictive abilities (Chapter I and Chapter II). This was the first empirical study to suggest that prediction is indeed an ability that is continuous across the lifespan.

Considering the theoretical implications of this dissertation, we have found that infants do experience and use prediction error, as operationalized by the Rescorla-Wagner model, suggesting infants have predictive abilities comparable to adults in this task (Chapter II). These results suggest that first, 6 month old infants have sufficient brain development to engage in prediction and second, prediction error is also a spark for infant learning, because infants learned to update their predictions according to modelling results. Taken together, these results suggest

that prediction is not just an ability that optimizes how adults behave in their everyday environment, but may also support learning from the very beginning of life.

Future studies should focus on the other subcomponents of prediction as prediction error is only a piece of the puzzle. However, using methods outside of pupillometry and designing a task to eliminate confounding variables is necessary for these investigations. For example, cognitive load has a clear link to pupil change (Ahern & Beatty, 1979; Alnæs et al., 2014; Borghini & Hazan, 2018; Hyönä, Tommola, & Alaja, 1995; for a review see van der Wel & van Steenbergen, 2018) and is a confounding factor when studying anticipation as the more an individual anticipates, the more likely they are to experience increasing cognitive load. In addition, studying the sensory representation of expectation, requires an infant friendly neuroimaging technique such as fNIRS (functional near-infrared spectroscopy). Lastly, two recent studies were published that examined infants' ability to update predictive models using different methods (*EEG*: Kayhan, Hunnius, O'Reilly & Bekkering, 2019; *pupillometry*: Kayhan, Meyer, O'Reilly, Hunnius & Bekkering, 2019) and found different results, which was attributed to differences in age, and task demands between the two studies. Clearly, more research is needed for characterizing prediction in infancy and importantly, future research should also focus on incorporating multiple subcomponents of prediction in the same study to understand how they work together.

As part of this work, we also learned a great deal about the limitations associated with pupillometry (Chapter III and Chapter V). Chapter III revealed pupillometry to not be compatible with visual statistical learning studies. Specifically, we showed that pupil size did not predict performance on the 2AFC task, which is administered at the end of the study to measure learning. In addition, we found pupil size to be highly influenced by baseline selection. Taken

together, we believe the fast-paced nature of VSL paradigms is not compatible with the slower nature of pupil change and future studies should ensure an adequate trial length, allowing for pupil size to fully evolve and return to baseline.

In Chapter V, we learned about the challenges of integrating eye movements and pupillometry in the same task. Despite finding eye movements that suggest infants and adults are making anticipations/expectations and replicating the results of Reuter, Emberson, Romberg & Lew-Williams (2018), pupil change did not reflect prediction error in this task. We believe the lack of pupil change findings are not caused by the influence of eye movements on pupil size because we counterbalanced where the target stimulus appeared for pre-switch and post-switch trials, and carefully selected a baseline, which allowed us to calculate pupil change from baseline rather than relying on the distorted raw pupil size. In addition, other studies have been successful in integrating eye movements and pupil size without these correction techniques (Cheng, Kaldy, Blaser 2019; Kleberg, del Bianco, Falck-Ytter, 2019; Ross-Sheehy & Eschman 2019), suggesting other factors are at play. One possible explanation for the pupil change outcome could be an inappropriate selection of time window for pupil change analysis, however, exploratory analyses suggest that to be unlikely. Thus, the more likely explanation is pupil change in this task might reflect a process other than prediction error, which explains the unexpected pupil results in the study.

Through these studies, we have learned a lot about task design that allows for optimal pupillometry measurements. First, to minimize pupil distortion caused by external factors, researchers should always have a baseline period and use isoluminant stimuli, which will prevent pupil size from changing as a result of brightness. Second, researchers must think carefully about the timing of their studies. We believe future pupillometry studies should have a sufficient period

of at least 3000 ms after stimulus onset to measure pupil change, as evidenced by previous successful infant studies that combined eye movements and pupil change (Cheng et al., 2019; Kleberg et al., 2019; Ross-Sheehy & Eschman, 2019; Zhang, Jaffe-Dax, Wilson, & Emberson, 2018). Overall, despite the challenges associated with pupillometry, there are solutions available that will make pupillometry a fruitful measure.

Regarding the null findings linking pupil dilation and prediction error in Chapter III, Chapter IV and Chapter V, this begs the question of how to interpret the positive findings linking pupil change and prediction error reported in Chapter II (and other studies reviewed in Chapter I). One possibility is that there is a potential file-drawer problem such that the link between pupil dilation and prediction error is over emphasized in publication. Future work would be needed to investigate if there is indeed a bias in publication towards positive results linking prediction error and pupillometry. Pursuing this line of thinking would involve conducting meta analyses to examine the p-values and effect sizes of established investigations and there are markers of publication bias and possibly pursuing strict replications of existing findings. Another possibility is that the complexity and differences in paradigms between studies is the cause of this discrepancy. For example, in another paradigm similar to the one used in Chapter II, involving data currently unpublished and not included as part of this dissertation, we also find evidence of pupil size reflecting prediction error (Zhang and Emberson, personal correspondence). In this paradigm, participants participated in one of three conditions: audio-audio association, visual-audio association, and visual-visual association. Similar to the task in Chapter II, after several learning trials, omission trials were introduced to elicit prediction error. All conditions presented the isoluminant visual stimuli in the center of the computer screen, and larger pupil dilation during omission trials were found for all these variations, albeit the pupil dilation for an auditory

omission is smaller than visual omission. These results suggest that under simple paradigm conditions involving centrally-presented stimuli and minimal eye movements, pupil dilation is a very clear measure of prediction error. While future work could be done to investigate whether there are publication biases towards using pupillometry to measure prediction error, it is also very likely that these divergent findings across studies are arising from the differences in the simplicity of the paradigms, with more paradigms, ranging from simple paradigms that rely solely on central fixation to more complex paradigms that require multiple measures and eye movements, needed to uncover this relationship.

Moving forward, it is equally important to focus on the benefits of using pupillometry. First, pupillometry allows for recording of the same physiological response across different ages in the same tasks (Ross-Sheehy & Eschman, 2019; Wetzel, Buttelmann, Schieler & Andreas 2016; Zhang et al., 2018). If we use similar tasks across populations, we can uncover the development of these cognitive processes. Second, pupillometry has a very high temporal resolution, which allows you to measure how cognitive processes unfold over time in as little as 2 ms intervals. Third, pupillometry can be used with other behavioral measures such as looking time and eye movements (Cheng et al., 2019; Kleberg et al., 2019; Ross-Sheehy & Eschman, 2019; Tummeltshammer, Feldman & Amso, 2019), providing new and important insights about infant cognition. In fact, pupillometry has already been used in the VOE literature (Jackson & Sirois, 2009; Laeng, Sirois, & Gredebäck, 2012; Sirois & Jackson, 2012) and short term memory (Johnson, Miller Singley, Peckham, Johnson & Bunge, 2014) and have proven to be useful in solving debates (Sirois & Jackson, 2012) and providing answers that would not be possible with other measures (Johnson et al., 2014). All of these cases demonstrate how pupillometry can be used to not only provide new insights about prediction error, but cognitive development in general.

Overall, although current research has uncovered a lot about what information infants can learn from, there is still much to uncover about how infants learn. Prediction is one possible learning mechanism, with numerous adult research indicating prediction error as the spark that initiates learning. Thus, as we continue to investigate how prediction and prediction error operates during infancy, the combination of pupillometry, eye movements and learning models, has helped and will continue to provide deeper insights about this topic. Future research should also continue to focus on examining the components of prediction separately and together in the same task. Together, these efforts will ultimately inform our understanding of how infants learn to make sense of the world.

References

Abla, D., Katahira, K., & Okanoya, K. (2008). On-line assessment of statistical learning by event-related potentials. *Journal of Cognitive Neuroscience*, *20*(6), 952–964.

Abla, D., & Okanoya, K. (2009). Visual statistical learning of shape sequences: An ERP study. *Neuroscience Research*, *64*(2), 185–190.

Addyman, C., Rocha, S., & Mareschal, D. (2014). Mapping the origins of time: Scalar errors in infant time estimation. *Developmental psychology, 50*(8), 2030.

Adler, S. A., & Haith, M. M. (2003). The nature of infants' visual expectations for event content. *Infancy*, *4*(3), 389-421.

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science, 205*(4412), 1289-1292.

Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., van de Pavert, S. H. P., & Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of Vision*, *14*(2014), 1–20. https://doi.org/10.1167/14.4.1

Arciuli, J., & Simpson, I. C. (2011). Statistical learning in typically developing children: the role of age and speed of stimulus presentation. *Developmental science*, *14*(3), 464-473.

Armstrong, B. C., Frost, R., & Christiansen, M. H. (2017). The long road of statistical learning research: Past, present and future. *Philosophical Transactions of the Royal Society B*, *372*(1711), 20160047.

Aslin, R. N. (1981) Development of smooth pursuit in human infants. In Fischer, D. F., Monty, R. A. & Senders, J. W. (Eds), *Eye movements: cognition and visual perception* (pp. 31-51). Hillsdale, NJ: Lawrence Erlbaum

Aslin, R. N. (2007). What's in a look?. *Developmental science*, *10*(1), 48-53.

Aslin, R. N. (2012). Infant eyes: A window on cognitive development. *Infancy*, *17*(1), 126-140.

Aslin, R. N. (2017). Statistical learning: a powerful mechanism that operates by mere exposure. *Wiley Interdisciplinary Reviews: Cognitive Science*, Vol. 8.

Aston-Jones, G., & Cohen, J. D. (2005). An Integrative Theory of Locus Function : Adaptive Gain and Optimal Performance. *Annual Review of Psychology*.

Atchison, D., & Smith, G. (2000). *Optics of the human eye*. Oxford: Butterworth-Heinemann.

Baillargeon, R. (1987). Object permanence in 3½-and 4½-month-old infants. *Developmental psychology, 23*(5), 655.

Baker, C. I., Olson, C. R., & Behrmann, M. (2004). Visual Statistical Learning Role of Attention and Perceptual Grouping in Visual Statistical Learning. *Psychological Science*, *15*(7), 1–8.

Bar, M. (2007). The proactive brain: using analogies and associations to generate predictions. *Trends in Cognitive Sciences*, *11*(7), 280–289. https://doi.org/10.1016/j.tics.2007.05.005

Batterink, L. J., & Paller, K. A. (2017). Online neural monitoring of statistical learning. *Cortex*, *90*.

Beesley, T., Nguyen, K. P., Pearson, D., & Le Pelley, M. E. (2015). Uncertainty and predictiveness determine attention to cues during human associative learning. *The Quarterly Journal of Experimental Psychology*, *68*(11), 2175-2199.

Benitez, V. L., & Saffran, J. R. (2018). Predictable events enhance word learning in toddlers. *Current Biology*, *28*(17), 2787-2793.

Benitez, V. L., & Smith, L. B. (2012). Predictable locations aid early object name learning. *Cognition, 125*(3), 339-352.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300.

Bergelson, E., & Swingley, D. (2012). At 6–9 months, human infants know the meanings of many common nouns. *Proceedings of the National Academy of Sciences*, *109*(9), 3253-3258.

Bertenthal, B. I., Boyer, T. W., & Harding, S. (2014). When do infants begin to follow a point?. *Developmental Psychology, 50*(8), 2036.

Borghini, G., & Hazan, V. (2018). Listening Effort During Sentence Processing Is Increased for Non-native Listeners: A Pupillometry Study. *Frontiers in Neuroscience*, *12*(March), 1–13. https://doi.org/10.3389/fnins.2018.00152

Brady, T. F., & Oliva, A. (2008). Statistical Learning Using Real-World Scenes Intent. *Psychological Science*, *19*(7), 678–685.

Braem, S., Coenen, E., Bombeke, K., van Bochove, M. E., & Notebaert, W. (2015). Open your eyes for prediction errors. *Cognitive, Affective, & Behavioral Neuroscience*, *15*(2), 374-380.

Browning, M., Behrens, T. E., Jocham, G., O'Reilly, J. X., & Bishop, S. J. (2015). Anxious individuals have difficulty learning the causal statistics of aversive environments. *Nature Neuroscience*, *18*(4), 590–596. https://doi.org/10.1038/nn.3961.Anxious

Bubic, A., Von Cramon, D. Y., & Schubotz, R. I. (2010). Prediction, cognition and the brain. *Frontiers in human neuroscience*, *4*, 25.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, *121*(1), 127–132.

Byers-Heinlein, K., Morin-Lessard, E., & Lew-Williams, C. (2017). Bilingual infants control their languages as they listen. *Proceedings of the National Academy of Sciences*.

Canfield, R. L., & Haith, M. M. (1991). Young infants' visual expectations for symmetric and asymmetric stimulus sequences. *Developmental Psychology*, *27*(2), 198.

Chen, Y. C., & Westermann, G. (2018). Different novelties revealed by infants' pupillary responses. *Scientific reports*, *8*(1), 1-8.

Cheng, C., Kaldy, Z., & Blaser, E. (2019). Focused attention predicts visual working memory performance in 13-month-old infants: A pupillometric study. *Developmental cognitive neuroscience*, *36*, 100616.

Clark, A. (2013). Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, 1–73.

Cohen, J. D., McClure, S. M., & Yu, A. J. (2007). Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *362*(1481), 933–942. https://doi.org/10.1098/rstb.2007.2098

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How Do Expectations Shape Perception? *Trends in Cognitive Sciences*, *22*(9), 764–779. https://doi.org/10.1016/j.tics.2018.06.002

DeLong, K. A., Urbach, T. P., & Kutas, M. (2005). Probabilistic word pre-activation during language comprehension inferred from electrical brain activity. *Nature neuroscience*, *8*(8), 1117-1121.

Den Ouden, H. E., Friston, K. J., Daw, N. D., McIntosh, A. R., & Stephan, K. E. (2009). A dual role for prediction error in associative learning. *Cerebral cortex*, *19*(5), 1175-1185.

Den Ouden, H. E., Kok, P., & De Lange, F. P. (2012). How prediction errors shape perception, attention, and motivation. *Frontiers in psychology*, *3*, 548.

Eckstein, M. K., Guerra-carrillo, B., Singley, A. T. M., & Bunge, S. A. (2016). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, 1–23.

Elsabbagh, M., Gliga, T., Pickles, A., Hudry, K., Charman, T., Johnson, M. H., & BASIS team. (2013). The development of face orienting mechanisms in infants at-risk for autism. *Behavioural brain research*, *251*, 147-154.

Emberson, L. L., Misyak, J. B., Schwade, J. A., Christiansen, M. H., & Goldstein, M. H. (2019). Comparing Statistical Learning Across Perceptual Modalities in Infancy: An Investigation of Underlying Learning Mechanism (s). *Developmental science*, e12847.

Emberson, L. L., Richards, J. E., & Aslin, R. N. (2015). Top-down modulation in the infant brain: Learning-induced expectations rapidly affect the sensory cortex at 6 months. *Proceedings of the National Academy of Sciences*, *112*(31), 9585-9590.

Evans, J. L., Saffran, J. R., & Robe-Torres, K. (2009). Statistical learning in children with specific language impairment. *Journal of Speech, Language, and Hearing Research*.

Falck-Ytter, T., Gredebäck, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature neuroscience*, *9*(7), 878.

Fawcett, C., Wesevich, V., & Gredebäck, G. (2016). Pupillary Contagion in Infancy Evidence for Spontaneous Transfer of Arousal. *Psychological science*, *27*(7), 997-1003.

Fiser, J., & Aslin, R. N. (2001). Unsupervised statistical learning of higher-order spatial structures from visual scenes. *Psychological science*, *12*(6), 499-504.

Franco, A., Gaillard, V., Cleeremans, A., & Destrebecqz, A. (2015). Assessing segmentation processes by click detection: online measure of statistical learning, or simple interference? *Behavior Research Methods*, *47*(4), 1393–1403.

Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, *360*(1456), 815-836.

Gagl, B., Hawelka, S., & Hutzler, F. (2011). Systematic influence of gaze position on pupil size measurement: analysis and correction. *Behavior research methods, 43*(4), 1171-1181.

Geangu, E., Hauf, P., Bhardwaj, R., & Bentz, W. (2011). Infant pupil diameter changes in response to others' positive and negative emotions. *PLoS ONE*, *6*(11). https://doi.org/10.1371/journal.pone.0027132

Gebhart, A. L., Aslin, R. N., & Newport, E. L. (2009). Changing Structures in Midstream: Learning Along the Statistical Garden Path. *Cognitive Science*, *33*(6), 1087–1116.

Geng, J. J., Blumenfeld, Z., Tyson, T. L., & Minzenberg, M. J. (2015). Pupil diameter reflects uncertainty in attentional selection during visual search. *Frontiers in Human Neuroscience*, (August), 1–14.

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience*, *10*(2), 252–269.

Goldinger, S. D., & Papesh, M. H. (2012). Pupil Dilation Reflects the Creation and Retrieval of Memories. *Current Directions in Psychological Science*, *21*(2), 90–95. https://doi.org/10.1177/0963721412436811

Gómez, D. M., Bion, R. A. H., & Mehler, J. (2011). The word segmentation process as revealed by click detection. *Language and Cognitive Processes*, *26*(2), 212–223.

Gredebäck, G., Lindskog, M., Juvrud, J. C., Green, D., & Marciszko, C. (2018). Action prediction allows hypothesis testing via internal forward models at 6 months of age. *Frontiers in psychology*, *9*, 290.

Gredebäck, G., & Melinder, A. M. D. (2010). Infants understanding of everyday social interactions: A dual process account. *Cognition, 114*, 197–206;

Gredebäck, G., von Hofsten, C., & Boudreau, J. P. (2002). Infants' visual tracking of continuous circular motion under conditions of occlusion and non-occlusion. *Infant Behavior and Development*, *25*(2), 161-182.

Háden, G. P., Németh, R., Török, M., & Winkler, I. (2015). Predictive processing of pitch trends in newborn infants. *Brain research*, *1626*, 14-20.

Haith, M. M., Hazan, C., & Goodman, G. S. (1988). Expectation and anticipation of dynamic visual events by 3.5-month-old babies. *Child development*, 467-479.

Hannon, E. E., & Trainor, L. J. (2007). Music acquisition: effects of enculturation and formal training on development. *Trends in cognitive sciences*, *11*(11), 466-472.

Hayes, T. R., & Petrov, A. A. (2015). Mapping and correcting the influence of gaze position on pupil size measurements. *Behavior Research Methods*.

Hepach, R., & Westermann, G. (2013). Infants' sensitivity to the congruence of others' emotions and actions. *Journal of experimental child psychology*, *115*(1), 16-29.

Hepach, R., & Westermann, G. (2016). Pupillometry in Infancy Research. *Journal of Cognition and Development*, (April). https://doi.org/10.1080/15248372.2015.1135801

Hespos, S. J., & Baillargeon, R. (2006). Décalage in infants' knowledge about occlusion and containment events: Converging evidence from action tasks. *Cognition, 99*(2), B31-B41.

Hespos, S., Gredebäck, G., Von Hofsten, C., & Spelke, E. S. (2009). Occlusion is hard: Comparing predictive reaching for visible and hidden objects in infants and adults. *Cognitive science*, *33*(8), 1483-1502.

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science, 143*(3611), 1190-1192.

Hogarth, L., Dickinson, A., Austin, A., Brown, C., & Duka, T. (2008). Attention and expectation in human predictive learning: The role of uncertainty. *The Quarterly Journal of Experimental Psychology*, *61*(11), 1658-1668.

Hollerman, J. R., & Schultz, W. (1998). Dopamine neurons report an error in the temporal prediction of reward during learning. *Nature neuroscience*, *1*(4), 304-309.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in infancy research*.

Hyönä, J., Tommola, J., & Alaja, A. M. (1995). Pupil Dilation as a Measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. *The Quarterly Journal of Experimental Psychology Section A*, *48*(3), 598–612.

Jackson, I., & Sirois, S. (2009). Infant cognition: going full factorial with pupil dilation. *Developmental science*, *12*(4), 670-679.

Jacobson, S. W., Jacobson, J. L., O'Neill, J. M., Padgett, R. J., Frankowski, J. J., & Bihun, J. T. (1992). Visual expectation and dimensions of infant information processing. *Child development*, *63*(3), 711-724.

Jin, K. S., Houston, J. L., Baillargeon, R., Groh, A. M., & Roisman, G. I. (2018). Young infants expect an unfamiliar adult to comfort a crying baby: Evidence from a standard violation-of-expectation task and a novel infant-triggered-video task. *Cognitive psychology*, *102*, 1-20.

Johnson, E. L., Miller Singley, A. T., Peckham, A. D., Johnson, S. L., & Bunge, S. A. (2014). Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Frontiers in psychology*, *5*, 218.

Kafkas, A., & Montaldi, D. (2011). Recognition memory strength is predicted by pupillary responses at encoding while fixation patterns distinguish recollection from familiarity. *The Quarterly Journal of Experimental Psychology*, *64*(10), 1971–1989.

Kafkas, A., & Montaldi, D. (2012). Familiarity and recollection produce distinct eye movement, pupil and medial temporal lobe responses when memory strength is matched. *Neuropsychologia*, *50*(13), 3080–3093.

Kafkas, A., & Montaldi, D. (2015). The pupillary response discriminates between subjective and objective familiarity and novelty. *Psychophysiology*, *52*(10), 1305–1316.

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*, *154*(3756), 1583–1585.

Kanakogi, Y., & Itakura, S. (2011). Developmental correspondence between action prediction and motor ability in early infancy. *Nature communications*, *2*(1), 1-6.

Kang, O. E., Huffer, K. E., & Wheatley, T. P. (2014). Pupil dilation dynamics track attention to high-level information. *PloS One*, *9*(8), e102463.

Kang, O. E., & Wheatley, T. P. (2015). Pupil dilation patterns reflect the contents of consciousness. *Consciousness and Cognition*, *35*(MAY 2015), 128–135.

Karuza, E. A., Emberson, L. L., & Aslin, R. N. (2014). Combining fMRI and behavioral measures to examine the process of human learning. *Neurobiology of Learning and Memory*, *109*, 193–206.

Karuza, E. A., Farmer, T. A., Alex, B., Smith, F. X., & Jaeger, T. F. (2014). On-line Measures of Prediction in a Self - Paced Statistical Learning Task. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *1*(36), 725–730.

Karuza, E. A., Kahn, A. E., Thompson-Schill, S. L., & Bassett, D. S. (2017). Process reveals structure: How a network is traversed mediates expectations about its architecture. *Scientific Reports*, *7*(1), 1–9.

Kayhan, E., Hunnius, S., O'Reilly, J. X., & Bekkering, H. (2019). Infants differentially update their internal models of a dynamic environment. *Cognition*, *186*, 139-146.

Kayhan, E., Meyer, M., O'Reilly, J. X., Hunnius, S., & Bekkering, H. (2019). Nine-month-old infants update their predictive models of a changing environment. *Developmental cognitive neuroscience*, *38*, 100680.

Kersten, D., & Yuille, A. (2003). Bayesian models of object perception. *Current opinion in neurobiology*, *13*(2), 150-158.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one, 7*(5), e36399.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2014). The Goldilocks effect in infant auditory attention. *Child development, 85*(5), 1795-1804.

Kim, R., Seitz, A., Feenstra, H., & Shams, L. (2009). Testing assumptions of statistical learning: Is it long-term and implicit? *Neuroscience Letters*, *461*(2), 145–149.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: Evidence for a domain general learning mechanism. *Cognition*, *83*(2), B35-B42.

Kirkham, N. Z., Slemmer, J. A., Richardson, D. C., & Johnson, S. P. (2007). Location, location, location: Development of spatiotemporal sequence learning in infancy. *Child development*, *78*(5), 1559-1571.

Kleberg, J. L., del Bianco, T., & Falck-Ytter, T. (2019). How Infants' Arousal Influences Their Visual Search. *Child development*, *90*(4), 1413-1423.

Kloosterman, N. A., Meindertsma, T., Loon, A. M., Lamme, V. A., Bonneh, Y. S., & Donner, T. H. (2015). Pupil size tracks perceptual content and surprise. *European Journal of Neuroscience*, *41*(8), 1068-1078.

Kochukhova, O., & Gredebäck, G. (2010). Preverbal infants anticipate that food will be brought to the mouth: an eye tracking study of manual feeding and flying spoons. *Child development*, *81*(6), 1729-1738.

Koenig, S., Uengoer, M., & Lachnit, H. (2018). Pupil dilation indicates the coding of past prediction errors: Evidence for attentional learning theory. *Psychophysiology*, *55*(4), 1–12.

Koelsch, S., Vuust, P., & Friston, K. (2019). Predictive processes and the peculiar case of music. *Trends in Cognitive Sciences*, *23*(1), 63-77.

Kouider, S., Long, B., Le Stanc, L., Charron, S., Fievet, A. C., Barbosa, L. S., & Gelskov, S. V. (2015). Neural dynamics of prediction and surprise in infants. *Nature communications*, *6*.

Kret, M. E., & Sjak-Shie, E. E. (2018). Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods*, (July 2018), 1–7. https://doi.org/10.3758/s13428-018-1075-y

Krogh, L., Vlach, H. A., & Johnson, S. P. (2013). Statistical Learning Across Development: Flexible Yet Constrained. *Frontiers in Psychology*, *3*(January), 1–11.

Kuipers, J. R., & Thierry, G. (2013). ERP-pupil size correlations reveal how bilingualism enhances cognitive flexibility. *Cortex*, *49*(10), 2853–2860. https://doi.org/10.1016/j.cortex.2013.01.012

Kutas, M., DeLong, K. A., & Smith, N. J. (2011). A look around at what lies ahead: Prediction and predictability in language processing. *Predictions in the brain: Using our past to generate a future*, 190-207.

LaBerge, D. (1995). *Attentional processing: The brain's art of mindfulness* (Vol. 2). Harvard University Press.

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious?. *Perspectives on psychological science*, *7*(1), 18-27.

Lagarias, J. C., Reeds, J. A., Wright, M. H., & Wright, P. E. (1998). Convergence properties of the Nelder--Mead simplex method in low dimensions. *SIAM Journal on optimization*, *9*(1), 112-147.

Lavin, C., San Martín, R., & Rosales Jubal, E. (2014). Pupil dilation signals uncertainty and surprise in a learning gambling task. *Frontiers in behavioral neuroscience*, *7*, 218.

Leung, E. H., & Rheingold, H. L. (1981). Development of pointing as a social gesture. *Developmental psychology, 17*(2), 215.

Lew-Williams, C., & Fernald, A. (2007). Young children learning Spanish make rapid use of grammatical gender in spoken word recognition. *Psychological Science*, *18*(3), 193-198.

Mathôt, S., van der Linden, L., Grainger, J., & Vitu, F. (2013). The pupillary light response reveals the focus of covert visual attention. *PloS One*, *8*(10), e78168. https://doi.org/10.1371/journal.pone.0078168

Mayo, J., & Eigsti, I. M. (2012). Brief report: A comparison of statistical learning in school-aged children with high functioning autism and typically developing peers. *Journal of autism and developmental disorders*, *42*(11), 2476-2485.

McClure, S. M., Berns, G. S., & Montague, P. R. (2003). Temporal prediction errors in a passive learning task activate human striatum. *Neuron*, *38*(2), 339-346.

Mill, R. D., O'Connor, A. R., & Dobbins, I. G. (2016). Pupil dilation during recognition memory: Isolating unexpected recognition from judgment uncertainty. *Cognition*, *154*, 81-94.

Misyak, J. B., Christiansen, M. H., & Tomblin, J. B. (2010). On-line individual differences in statistical learning predict language processing. *Frontiers in psychology*, *1*, 31.

Morissette, P., Ricard, M., & Décarie, TG (1995). Joint visual attention and pointing in infancy: A longitudinal study of comprehension. *British Journal of Developmental Psychology, 13* (2), 163-175.

Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'connell, R. G. (2011). Pupillometry and P3 index the locus coeruleus-noradrenergic arousal function in humans. *Psychophysiology*, *48*(11), 1532–1543.

Murphy, P. R., Van Moort, M. L., & Nieuwenhuis, S. (2016). The pupillary orienting response predicts adaptive behavioral adjustment after errors. *PLoS ONE*, *11*(3).

Nagy, Z., Westerberg, H., & Klingberg, T. (2004). Maturation of white matter is associated with the development of cognitive functions during childhood. *Journal of cognitive neuroscience*, *16*(7), 1227-1233.

Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasly, B., & Gold, J. I. (2012). Rational regulation of learning dynamics by pupil-linked arousal systems. *Nature neuroscience*, *15*(7), 1040-1046.

O'Reilly, J. X., Schüffelgen, U., Cuell, S. F., Behrens, T. E., Mars, R. B., & Rushworth, M. F. (2013). Dissociable effects of surprise and model update in parietal and anterior cingulate cortex. *Proceedings of the National Academy of Sciences*, *110*(38), E3660-E3669.

Pätzold, W., & Liszkowski, U. (2019). Pupillometry reveals communication-induced object expectations in 12- but not 8-months-old infants. *Developmental Science*. https://doi.org/10.1111/desc.12832

Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological review, 87*(6), 532.

Pearce, J. M., Kaye, H., & Hall, G. (1982). Predictive accuracy and stimulus associability: Development of a model for Pavlovian learning. In M. Commonsa, R. Herrnstein, & A. R. Wagner (Eds.), *Quantitative analysis of behavior* (pp. 241–255). Cam- bridge, MA: Ballinger.

Pearce, J. M., & Mackintosh, N. J. (2010). Two theories of attention: A review and a possible integration. In C. Mitchell & M. E. Le Pelley, *Attention and learning* (pp. 11–39). Oxford, UK: Oxford University Press.

Preuschoff, K., Marius't Hart, B., & Einhäuser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in neuroscience*, *5*.

Raisig, S., Welke, T., Hagendorf, H., & van der Meer, E. (2010). I spy with my little eye: Detection of temporal violations in event sequences and the pupillary response. *International Journal of Psychophysiology*, *76*(1), 1-8.

Rao, R. P. N., & Ballard, D. H. (1995). Natural basis functions and topographic memory for face recognition. *Proceedings of the International Joint Conferencve on Artificial Intelligence (IJCAI)*, 10–17.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79-87

Raviv, L., & Arnon, I. (2017). The developmental trajectory of children's auditory and visual statistical learning abilities: modality-based differences in the effect of age. *Developmental Science*.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical conditioning II: Current research and theory*, *2*, 64-99.

Reznick, J. S., Chawarska, K., & Betts, S. (2000). The development of visual expectations in the first year. *Child Development, 71*(5), 1191-1204.

Reuter, T., Emberson, L., Romberg, A., & Lew-Williams, C. (2018). Individual differences in nonverbal prediction and vocabulary size in infancy. *Cognition*, *176*, 215-219.

Romberg, A. R., & Saffran, J. R. (2013). All together now: Concurrent learning of multiple structures in an artificial language. *Cognitive science*, *37*(7), 1290-1320.

Rose, S. A., Feldman, J. F., Jankowski, J. J., & Caro, D. M. (2002). A longitudinal study of visual expectation and reaction time in the first year of life. *Child Development*, *73*(1), 47-61.

Ross-Sheehy, S., & Eschman, B. (2019). Assessing visual STM in infants and adults: eye movements and pupil dynamics reflect memory maintenance. *Visual Cognition*, *27*(1), 78-92.

Ruffman, T., Slade, L., & Redman, J. (2005). Young infants' expectations about hidden objects. *Cognition, 97*(2), B35-B43.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926-1928.

Saffran, J. R., Johnson, E. K., Aslin, R. N., & Newport, E. L. (1999). Statistical learning of tone sequences by human infants and adults. *Cognition*, *70*(1), 27-52.

Saffran, J. R., & Kirkham, N. Z. (2018). Infant Statistical Learning. *Annual Review of Psychology*, *69*(1).

Saffran, J. R., Newport, E. L., Aslin, R. N., Tunick, R. A., & Barrueco, S. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological science*, *8*(2), 101-105.

Samuels, E. R., & Szabadi, E. (2008). Functional neuroanatomy of the noradrenergic locus coeruleus: its roles in the regulation of arousal and autonomic function part II: physiological and pharmacological manipulations and pathological alterations of locus coeruleus activity in humans. *Current neuropharmacology, 6*(3), 254-285.

Scheepers, C., Mohr, S., Fischer, M. H., & Roberts, A. M. (2013). Listening to limericks: a pupillometry investigation of perceivers' expectancy. *PloS one*, *8*(9), e74986.

Siegelman, N., Bogaerts, L., Christiansen, M. H., & Frost, R. (2017). Towards a theory of individual differences in statistical learning. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *372*(1711), 20160059.

Siegelman, N., Bogaerts, L., Kronenfeld, O., & Frost, R. (2018). Redefining "Learning" in Statistical Learning: What Does an Online Measure Reveal About the Assimilation of Visual Regularities? *Cognitive Science*, *42*(1996), 692–727.

Sirois, S., & Jackson, I. R. (2012). Pupil dilation and object permanence in infants. *Infancy*, *17*(1), 61-78.

Smith, L. B., & Yu, C. (2013). Visual attention is not enough: Individual differences in statistical word-referent learning in infants. *Language Learning and Development*, *9*(1), 25-49.

Smyser, C. D., Inder, T. E., Shimony, J. S., Hill, J. E., Degnan, A. J., Snyder, A. Z., & Neil, J. J. (2010). Longitudinal analysis of neural net work development in preterm infants. *Cerebral cortex*, *20*(12), 2852-2862.

Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, *348*(6230), 91-94.

Sterpenich, V., D'Argembeau, A., Desseilles, M., Balteau, E., Albouy, G., Vandewalle, G., ... & Maquet, P. (2006). The locus ceruleus is involved in the successful retrieval of emotional memories in humans. *Journal of Neuroscience, 26*(28), 7416-7423.

Summerfield, C., & de Lange, F. P. (2014). Expectation in perceptual decision making: neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756. https://doi.org/10.1038/nrn3838

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409. https://doi.org/10.1016/j.tics.2009.06.003

Tamási, K., McKean, C., Gafos, A., Fritzsche, T., & Höhle, B. (2017). Pupillometry registers toddlers' sensitivity to degrees of mispronunciation. *Journal of experimental child psychology*, *153*, 140-148.

Trainor, L. J., & Trehub, S. E. (1992). A comparison of infants' and adults' sensitivity to Western musical structure. *Journal of Experimental Psychology: Human Perception and Performance*, *18*(2), 394.

Tsukahara, J. S., Harrison, T. L., & Engle, R. W. (2016). The relationship between baseline pupil size and intelligence. *Cognitive Psychology*, *91*.

Tummeltshammer, K., Amso, D., French, R. M., & Kirkham, N. Z. (2017). Across space and time: Infants learn from backward and forward visual statistics. *Developmental science, 20*(5), e12474.

Tummeltshammer, K., Feldman, E. C., & Amso, D. (2019). Using pupil dilation, eye-blink rate, and the value of mother to investigate reward learning mechanisms in infancy. *Developmental cognitive neuroscience*, *36*, 100608.

Turk-Browne, N. B., Isola, P. J., Scholl, B. J., & Treat, T. A. (2008). Multidimensional visual statistical learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*(2), 399–407.

Turk-Browne, N. B., & Scholl, B. J. (2009). Flexible visual statistical learning: Transfer across space and time. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(1), 195–202.

Turk-Browne, N. B., Scholl, B. J., Chun, M. M., & Johnson, M. K. (2010). Neural Evidence of Statistical Learning: Efficient Detection of Visual Regularities Without Awareness. *Journal of Cognitive Neuroscience*, *21*(10), 1934–1945.

Turk-Browne, N. B., Scholl, B. J., Johnson, M. K., & Chun, M. M. (2010). Implicit perceptual anticipation triggered by statistical learning. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *30*(33), 11177–11187.

Turk -Browne, N. B., Jungé, J., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology. General*, *134*(4), 552–564.

Unsworth, N., & Robison, M. K. (2017). The importance of arousal for variation in working memory capacity and attention control: A latent variable pupillometry study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(12), 1962.

Van Berkum, J. J., Brown, C. M., Zwitserlood, P., Kooijman, V., & Hagoort, P. (2005). Anticipating upcoming words in discourse: evidence from ERPs and reading times. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(3), 443.

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin and Review*, 1–11. https://doi.org/10.3758/s13423-018-1432-y

Walden, T., Kim, G., McCoy, C., & Karrass, J. (2007). Do you believe in magic? Infants' social looking during violations of expectations. Developmental science, 10(5), 654-663.

Wang, S., Baillargeon, R., & Brueckner, L. (2004). Young infants ' reasoning about hidden objects : evidence from violation-of-expectation tasks with test trials only. *Cognition*, *93*, 167–198. https://doi.org/10.1016/j.cognition.2003.09.012

Wei, X. X., & Stocker, A. A. (2015). A Bayesian observer model constrained by efficient coding can explain anti-Bayesian percepts. *Nature neuroscience, 18*(10), 1509.

Wells, J. B., Christiansen, M. H., Race, D. S., Acheson, D. J., & MacDonald, M. C. (2009). Experience and sentence processing: Statistical learning and relative clause comprehension. *Cognitive psychology*, *58*(2), 250-271.

Wentworth, N. (2009). Future Orientation. *Encyclopedia of Infant and Early Childhood Development*, (2005), 562–574. https://doi.org/10.1007/b106810

Wentworth, N., & Haith, M. M. (1992). Event-specific expectations of 2-and 3-month-old infants. *Developmental Psychology*, *28*(5), 842.

Wentworth, N., Haith, M. M., & Hood, R. (2002). Spatiotemporal regularity and interevent contingencies as information for infants' visual expectations. *Infancy*, *3*(3), 303-321.

Wetzel, N., Buttlemann, D., Schieler, A., & Widmann, A. (2015). Infant and adult pupil dilation in response to unexpected sounds. *Developmental psychobiology, 58*(3), 382-392.

Wills, A. J., Lavric, A., Croft, G. S., & Hodgson, T. L. (2007). Predictive learning, prediction errors, and attention: Evidence from event-related potentials and eye tracking. *Journal of Cognitive Neuroscience*, *19*(5), 843-854.

Winn, B., Whitaker, D., Elliott, D. B., & Phillips, N. J. (1994). Factors affecting light-adapted pupil size in normal human subjects. *Investigative ophthalmology & visual science, 35*(3), 1132-1137.

Wong-Kee-You, A. M., & Adler, S. A. (2016). Anticipatory eye movements and long-term memory in early infancy. *Developmental psychobiology*, *58*(7), 841-851.

Wykowska, A., Anderl, C., Schubö, A., & Hommel, B. (2013). Motivation modulates visual attention: evidence from pupillometry. *Frontiers in Psychology*, *4*(February), 59. https://doi.org/10.3389/fpsyg.2013.00059

Yechiam, E., & Telpaz, A. (2011). To take risk is to face loss: A tonic pupillometry study. *Frontiers in Psychology*, *2*(NOV), 1–9. https://doi.org/10.3389/fpsyg.2011.00344

Ylinen, S., Bosseler, A., Junttila, K., & Huotilainen, M. (2017). Predictive coding accelerates word recognition and learning in the early stages of language development. *Developmental science, 20*(6), e12472.

Yu, C., & Smith, L. B. (2007). Rapid word learning under uncertainty via cross-situational statistics. *Psychological science*, *18*(5), 414-420.

Zhang, F., & Emberson, L. L. (2019). Opposing timing constraints severely limit the use of pupillometry to investigate visual statistical learning. *Frontiers in psychology, 10,* 1792.

Zhang, F., Jaffe-Dax, S., Wilson, R., & Emberson, L. (2018). Prediction in infants and adults: A pupillometry study. *Developmental Science*, (October 2017), e12780.

Appendix A

Additional Methods Information For Chapter II

Stimuli

We used an isoluminance script provided by Daniel Swingley (personal communication, David Brainard and Daniel Swingley, 2016) which takes a standard RGB image, corrects the gamma, separates the image's chromaticity and luminosity, and sets the luminance to a constant value, thereby making the image isoluminant, without affecting the chromaticity. We selected a luminance value of 0.3 to match the lighting in the room and not distort the visual stimuli and videos to the point of being unrecognizable.

Design and Procedure

The monitor measured 34cm by 27cm, and eye gaze was recorded using the 25mm infant lens. Adult gaze was recorded at 500Hz whereas infant gaze was recorded at 250Hz (some were recorded at 500Hz but this data was downsampled to 250Hz by removing every other row of data starting at 0ms). Luminance in the room was measured using Light Meter, a mobile app used to measure light intensity (version 1.1, Elena Polyanskaya). The display monitor (Eyelink 1000 eye tracker) was facing the participant. The host monitor plus experimenting computer were in front of the experimenter. Before beginning the experiment, a five-point calibration was used. We performed calibration and validation for all participants but did not exclude participants based on validation accuracy[7].

[7] Validation accuracy does not influence the accuracy of pupil size measurement (personal communication, Dan McEchron, 2016), especially given that our stimuli were all presented centrally.

Infants sat on their caregivers' laps throughout the experiment. Caregivers were instructed to not interfere with the infant and wore a visor during the experiment, which prevented them from seeing the screen and blinding them to the content of the individual trials (i.e., to prevent biasing of infant behaviour). The experimenter watched the caregiver and infant from the Eyelink 1000 host computer in order to exit the experiment when infants became too inattentive and fussy. The experimenter adjusted the display monitor as infants moved.

Preprocessing

Data files exported from the eye tracker were processed to determine pupil measurements for each participant. First, missing data of up to 85 samples (i.e. blinks) for adults were interpolated linearly (Hepach & Westermann, 2013; Addyman, Rocha & Mareschal, 2014; Fawcett, Wesevich & Gredebäck, 2016) using the package zoo (Version 1.7-13, Zeileis, Grothendieck, Ryan & Andrews) for the R software (Version 3.2.1; R Development Core Team, 2015). We skipped this step for infants because there was no way to confirm whether the missing sample was a result of a blink or infants looking away. Next, trials with more than 50% of data missing (Fawcett et al., 2016) during the first 4000ms were removed from the analyses (44% or 808 trials for infants, and 1.2% or 45 trials for adults). The first 4000ms of a trial captures the predicting auditory stimulus (1000ms), the predicted visual stimulus (1000ms), and approximately 2000ms to allow for pupil response to unfold. This resulted in average number of 8.9 omission (range 1 – 22) and 27.2 present trials (range 6 – 58) for infants. We excluded a total of 4 infants because they had 0 omission or present trials. Adults had an average of 33.6 omission (range 30 – 34) 79.0 present trials (range 72 – 80).

Pupil size was recorded with pupil area instead of pupil diameter and converted to pupil diameter with an equation diameter = 256 * sqrt(area / Pi). The adult data was smoothed with a moving average over 50 samples or 100ms to remove outliers (Fawcett et al., 2016). We took a different approach to remove outliers for infants because some infants had extreme raw pupil size values (i.e. numbers that were extremely outside what a normal pupil size should be) due to movement and interpolating based on those numbers was found to increase the variability of the data. We visually examined the raw pupil data of infants and set a liberal minimum and maximum pupil size threshold of 2000 and 9000 arbitrary units respectively (SR Research, 2013, p. 120). The remaining missing data was linearly interpolated. All time points are calculated relative to the onset of the predicting sound.

Lastly, we calculated the PDR as percentage change from baseline. We included a video at the end of each trial because we wanted to calculate baseline by taking 500ms before stimulus onset (Cavanagh, Wiecki, Kochar & Frank, 2014; Unsworth & Robison, 2014; Hochmann & Papeo, 2014); however, the end of the video was also a period of time where infants would often look away. Therefore, we changed our baseline to the period after the predicting sound but before the predicted visual stimulus appears, for a total duration of 1000ms. This period was less noisy due to high attentiveness and looking compliance, making it was more suitable for a standard baseline period. This baseline measure was subtracted from each pupil diameter measurement and divided by baseline and then multiplied by one hundred to get the PDR.

Modeling PDR Response Using A Learning Model

The parameters that were used to fit the model to the PDR are the initial prediction for the appearance of visual stimuli ($P(0)$) and the initial learning rate ($\alpha(0)$). These parameters were

optimized using Nelder-Mead simplex method (Lagarias et al., 1998). The parameters were restricted to be non-negative and to be not greater than 1. For each trial of every subject the fitted model estimated the prediction error ($\delta(t)$) and prediction ($P(t)$). For each participant, we initialized the fit from 625 evenly spaced starting points and used the lowest achieved MSE as the best fit.

Appendix B

Experiment 2: Pupil Change Across Triplet Presentation (Online Measure of Learning) Using Unbaselined Pupil Diameter

We ran the same GLMMs for Experiment 2, using unbaselined pupil diameter, for the anticipatory and viewing period. For the anticipatory period, we found a familiarity effect (β = -7.8, $p < 0.001$) but no effect of image position (β = -1.2, $p = 0.81$) and accuracy (β = -552.7, $p = 0.58$). For the viewing period, we found a familiarity effect (β = -7.2, $p < 0.001$) but no effect of image position (β = 1.2, $p = 0.79$) and accuracy (β = -555.5, $p = 0.58$). Overall, we find that unbaselined pupil size during the entire trial decreased as a result of familiarity, which could be reflective of diminishing uncertainty. However, this decrease in unbaselined pupil size is not predictive of accuracy, so it is unlikely that this change is related to learning effects.

Appendix C

Experiment 2: Siegelman et al. (2018) Online Measure of SL Using Unbaselined Pupil Diameter

When we calculate Siegelman et al. (2018) measure of SL using unbaselined pupil diameter, we find that the trajectory of the online measure was not significant for Experiment 2 for both the anticipatory ($r = -0.15$, $df = 22$, $p = 0.48$) and viewing period ($r = -0.34$, $df = 22$, $p = 0.1$). Furthermore, results revealed a non-significant relationship between Siegelman et al. (2018)'s online measure of SL and accuracy on test phase (anticipatory period: $r = 0.21$, $df = 47$, $p = 0.14$, viewing period: $r = -0.015$, $df = 47$, $p = 0.92$). These results also indicate that unbaselined pupil diameter is also not an online measure for learning in VSL.

Appendix D

Experiment 2: Pupil Change Across Triplet Presentation (Online Measure of Learning) Using Second Half of Familiarization Trials

We ran the same GLMMs for Experiment 2, using the second half of trials, for the anticipatory and viewing period. For the anticipatory period, we found that the simplest model, with only image position as a predictor result in the best GLMM model. The model revealed that image position ($\beta = 9.5\text{e-}3, p < 0.001$) had a significant effect on pupil change. For the viewing period, we found a significant familiarity effect ($\beta = 1.4\text{e-}3, p < 0.05$), marginal image position effect ($\beta = 4.3\text{e-}3, p = 0.1$) but no effect of accuracy ($\beta = 4.2\text{e-}3, p = 0.94$).

Appendix E

Experiment 2: Siegelman et al. (2018) Online Measure of SL Using Second Half of Familiarization Trials

We calculated Siegelman et al. (2018) measure of SL using the second half of trials. Results showed the trajectory of the online measure was not significant for Experiment 2 for both the anticipatory ($r = -0.19, p = 0.56$) and viewing period ($r = 0.09, p = 0.52$). Furthermore, results revealed a non-significant relationship between Siegelman et al. (2018)'s online measure of SL and accuracy on test phase (anticipatory period: $r = 0.16, p = 0.28$, viewing period: $r = -0.08, p = 0.58$). Overall, we find that by investigating the second half of the experiments, where we can be fairly confident participants have learned, accuracy is still not predictive of pupil change and pupil change is not reflective of online learning.

Appendix F

Experiment 2: Pupil Change Across Triplet Presentation (Online Measure of Learning) Using Participants that Performed Above Median Split

We reanalyzed the data from Experiment 2, by focusing on the participants who performed above median (60%). This group of participants has a median accuracy of 74%. We ran the same GLMMs as Experiment 2. For the anticipatory period, we found that the simplest model, with only image position as a predictor result in the best GLMM model. The model revealed that image position (β = 6.2e-3, $p < 0.01$) had a significant effect on pupil change. For the viewing period, we found a significant effect of familiarity (β = 1.3e-3, $p < 0.001$) but no effect of accuracy (β = -6.7e-3, $p = 0.59$) or image position (β = 2.4e-3, $p = 0.33$). These results confirm that it is not the inclusion of poor learners (as measured by post-test performance) that obfuscates the pupillometry effects.

Appendix G

Experiment 2: Siegelman et al. (2018) Online Measure of SL Using Participants that Performed Above Median Split

We also calculated Siegelman et al. (2018) measure of SL. Results showed the trajectory of the online measure was not significant for Experiment 2 for both the anticipatory ($r = -0.22$, $p = 0.29$) and viewing period ($r = -0.24$, $p = 0.26$). Furthermore, results revealed a non-significant relationship between Siegelman et al. (2018)'s online measure of SL and accuracy on test phase (anticipatory period: $r = 0.36$, $p = 0.1$, viewing period: $r = 0.15$, $p = 0.51$). Overall, we find that by investigating the good learners, with a median accuracy of 74%, pupil change is still not a sensitive measure of online learning.

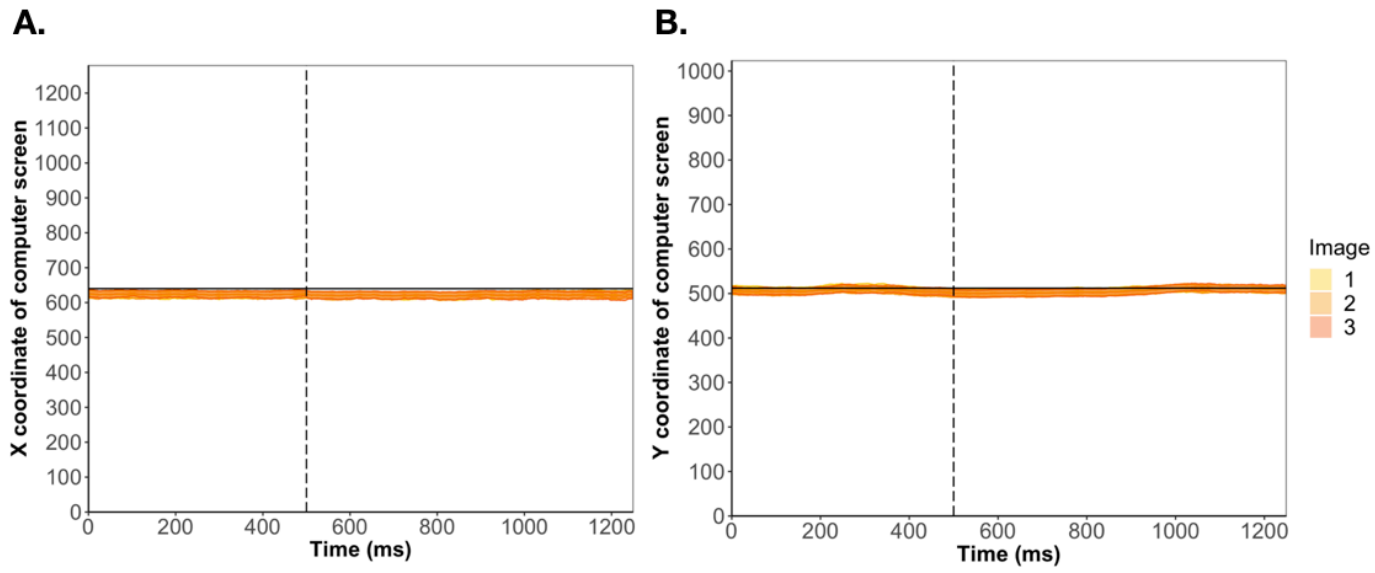
Appendix H

Average Gaze Position For Trial In Experiment 2

A. Plot shows where along the horizontal dimension of the screen participants were looking. The horizontal black line in both plots indicates the center of the screen, where the visual stimulus appeared.

B. The second plot shows where along the vertical dimension of the screen participants were looked.

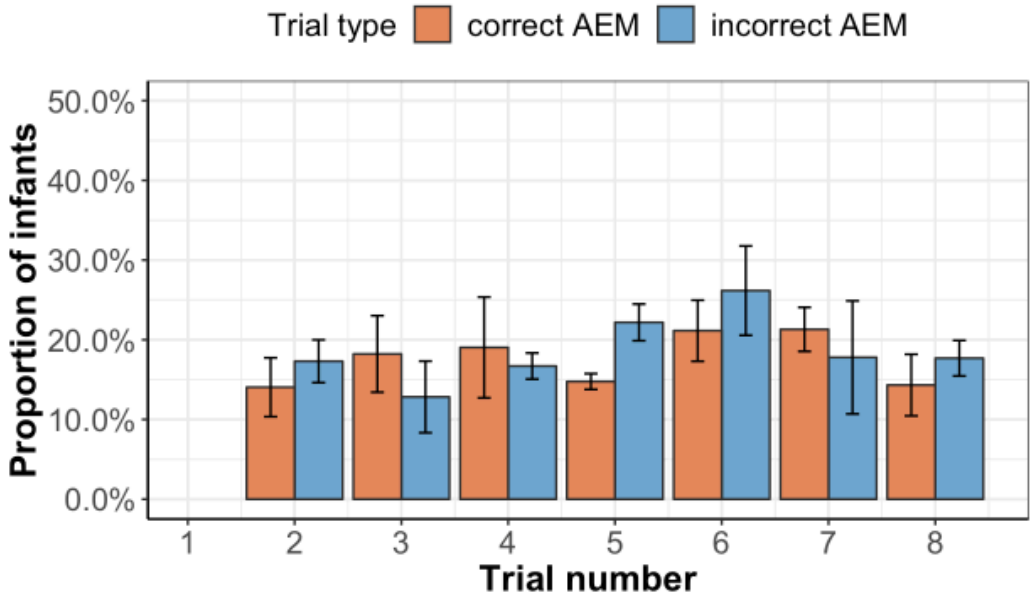
Taken together, the plots show that participants were looking at the center of the screen during the entire trial, confirming changes in pupil change is not a result of eye movement and participants were attending to the visual stimuli.



Appendix I

Proportion Of Infants That Generated A Correct AEM Versus Incorrect AEM On A Trial By Trial Basis During Pre-Switch.

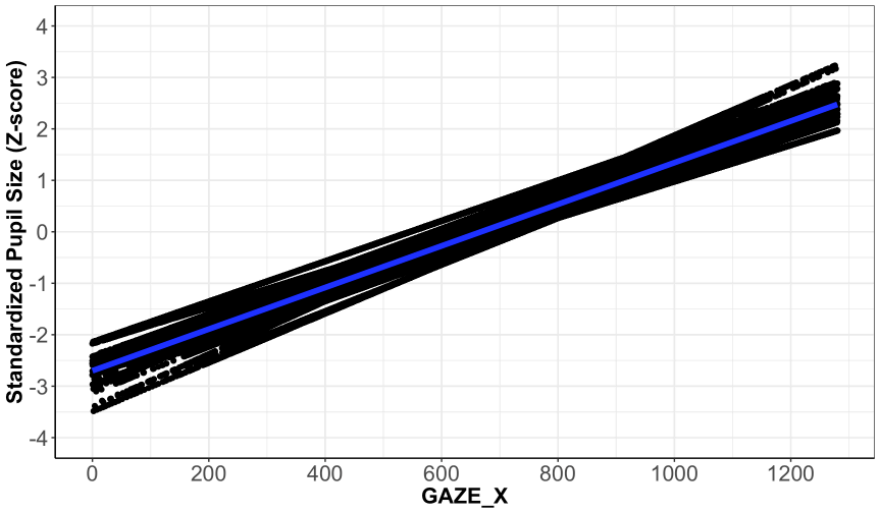
Vertical error bars indicate between block SEM. We conducted a linear regression using trial number in pre-switch and trial type as the independent variables and proportion of infants as the dependent variable. We found no significant effect of trial number ($B = 0.0033, p = 0.67$) or trial type ($B = -0.0092, p = 0.88$) and no significant interaction ($B = 0.0041, p = 0.71$).



Appendix J

Changes In Pupil Size, Z-Scored To Standardize Pupil Size For Infants, As A Function Of X Gaze Coordinates (I.E. Width Of The Screen).

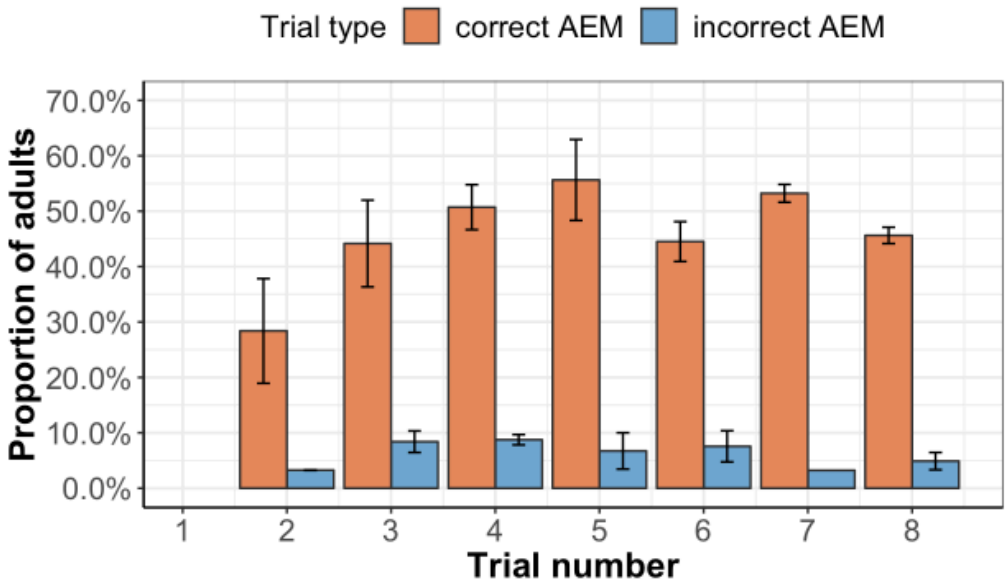
As infants' eyes move from left to right, pupil size increases.



Appendix K

Proportion Of Adults That Generated A Correct AEM Versus Incorrect AEM On A Trial By Trial Basis During Pre-Switch.

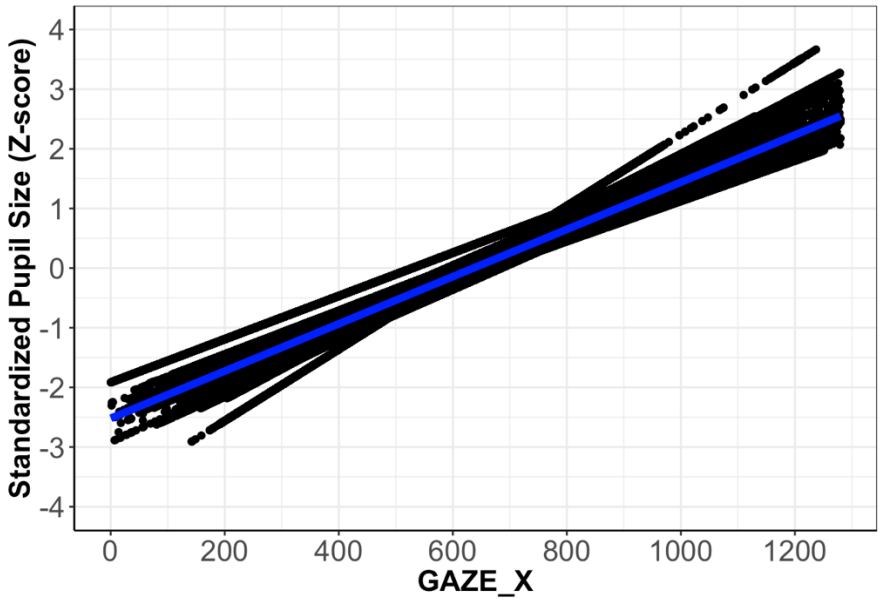
Vertical error bars indicate between block SEM. We conducted a linear regression using trial number in pre-switch and trial type as the independent variables and proportion of adults as the dependent variable. We found a significant effect of trial number ($B = 0.023, p = 0.025$) and trial type ($B = -0.28, p < 0.001$) but no significant interaction ($B = -0.023, p = 0.14$).



Appendix L

Changes In Pupil Size, Z-Scored To Standardize Pupil Size For Adults, As A Function Of X Gaze Coordinates (I.E. Width Of The Screen).

As adults' eyes move from left to right, pupil size increases.



Appendix M

Poster References

Zhang, F., Reuter, T. & Emberson, L. L. (2019, October). Prediction and prediction error in 14-month-old infants. Cognitive Development Society, Louisville, Kentucky.

Zhang, F. & Emberson, L. L. (2019, March). Prediction Error in Auditory and Visual Domain of 6-month-old: A Pupillometry Study. Society for Research in Child Development 2019 Biennial Meeting, Baltimore, Maryland.

Zhang, F. & Emberson, L. L. (2018, July). Prediction Error in Auditory and Visual Domain of 6-month-old: A Pupillometry Study. International Congress of Infant Society Biennial Congress, Philadelphia, Pennsylvania.

Zhang, F., Jaffe-Dax, S., Wilson, R. C. & Emberson, L. L. (2018, July). Prediction in Infants and Adults: A Pupillometry Study. International Congress of Infant Society Biennial Congress, Philadelphia, Pennsylvania.

Zhang, F. & Emberson, L. L. (2018, June). Prediction Error in Auditory and Visual Domain of 6-month-old: A Pupillometry Study. 1st International Workshop on Predictive Processing, San Sebastián, Spain.

Zhang, F., Jaffe-Dax, S., Wilson, R. C. & Emberson, L. L. (2018, June). Prediction in Infants and Adults: A Pupillometry Study. 1st International Workshop on Predictive Processing, San Sebastián, Spain.

Zhang, F., Jaffe-Dax, S., Wilson, R. C. & Emberson, L. L. (2017, June). Prediction and statistical learning in infants and adults: A Pupillometry Study. 1st International Conference on Interdisciplinary Advances in Statistical Learning, Bilbao, Spain.

Zhang, F. & Emberson, L. L. (2017, June). Examining the online mechanisms of visual statistical learning using pupillometry. International Conference on Interdisciplinary Advances in Statistical Learning, Bilbao, Spain.

Zhang, F., Jaffe-Dax, S., Wilson, R. C. & Emberson, L. L. (2017, April). Prediction in young infants: A pupillometry study. Society for Research in Child Development 2017 Biennial Meeting, Austin, Texas.