

In press, *Infancy*

Does surprise enhance infant memory? Assessing the impact of the encoding context on subsequent object recognition

Csink, Viktoria^a, Mareschal, Denis^b, Gliga, Teodora^c

^a Birkbeck University of London, Department of Psychological Sciences, Malet St, Bloomsbury, London WC1E 7HX, United Kingdom, viktoria.csink@gmail.com, +44(0)7597769271

^b Birkbeck University of London, Department of Psychological Sciences, Malet St, Bloomsbury, London WC1E 7HX, United Kingdom, d.mareschal@bbk.ac.uk, +44 (0)20 7631 6582

^c University of East Anglia, School of Psychology, Norwich Research Park, Norwich NR4 7TJ, United Kingdom, T.Gliga@uea.ac.uk, +44 (0)1603 59 7967

Corresponding author: **Csink, Viktoria**

Keywords: surprise, recognition memory, pupillometry, mislabelling, enhanced learning

Acknowledgements

The research was funded by the Centre for Brain and Cognitive Development, Birkbeck University of London. The authors declare no conflicts of interest with regard to the funding source for this study.

Abstract

A discrepancy between what was predicted and what is observed has been linked to increased looking times, changes in brain electrical activity, and increased pupil dilation in infants. These processes associated with heightened attention and readiness to learn might enhance the encoding and memory consolidation of the surprising object, as suggested by both the infant and the adult literature. We therefore investigated whether the presence of surprise during the encoding context enhances subsequent encoding and recognition memory processes for the items that violated infants' expectations. Seventeen-month-olds viewed 20 familiar objects, half of which were labelled correctly, while the other half were mislabelled. Subsequently, infants were presented with a silent recognition memory test where the previously labelled objects appeared along with new images. Pupil dilation was measured, with more dilated pupils indicating (1) surprise during those labelling events where the item was mislabelled and (2) successful retrieval processes during the memory test. Infants responded with more pupil dilation to mislabelling compared to correct labelling. Importantly, despite the presence of a surprise response during mislabelling, infants only differentiated between the previously seen and unseen items at the memory test, offering no evidence that surprise had facilitated the encoding of the mislabelled items.

Keywords: surprise, recognition memory, pupillometry, mislabelling, enhanced learning.

Introduction

The idea that infants are active learners who systematically select information for further exploration and learning has gained support from studies showing that infants preferentially attend to novel stimuli as a function of its complexity (Aslin, 2007) and that they learn better about objects that they themselves have chosen as a source of further information (Begus, Gliga & Southgate, 2014). In addition, infants have been shown to increase their sampling of visual and auditory stimuli that have an intermediate level of complexity, reducing their exploration to both minimally and maximally complex input (Kidd, Piantadosi & Aslin, 2012; 2014). In light of infants' limited capacity to process new information, stimulus novelty and the relative uncertainty of event sequences might drive their allocation of attention to stimuli that is neither perfectly predictable, nor impossible to learn from.

In accordance with the idea that a certain degree of complexity enhances learning, a categorization study showed that infants learnt best when the perceptual distance across consecutive exemplars was increased (Mather & Plunkett, 2011). A computational model using the same stimuli set has shown further learning benefits when the model chose successive stimuli that maximised the feature distance between the new stimulus and the previously encoded one (Twomey & Westermann, 2018). Although larger information gaps temporarily increase the uncertainty in the environment, they might facilitate learning by triggering cognitive processes that are directed at closing these gaps.

Violations of prior predictions might thus play a special role in learning by highlighting a new piece of information and facilitating the revision of the learner's current set of expectations. Early studies in developmental research have shown that both infants (Charlesworth, 1966) and primary school children (Charlesworth, 1964) continued an experiment longer and completed more trials in the conditions where their prior predictions were overwritten by unexpected events. In addition, pre-schoolers have been found to increase

their exploration of toys that have previously violated their expectations (Bonawitz et al., 2012; Cook, Goodman & Schulz, 2011), possibly indicating that longer looking times in the traditional violation of expectation paradigms (Baillargeon, 1986; Baillargeon, Spelke & Wasserman, 1985; Wynn, 1992) reflect infants' attempt to visually explore the event outcome that contradicted their expectations. Furthermore, the magnitude of surprise measured by pupil dilation has been linked to the strength of infants' prior predictions (Gredebäck et al., 2018; Juvrud et al., 2019), indicating that the mismatch between infants' expectations and the event outcome creates an error term, which is then fed back into the cognitive system.

Violations of strong pre-existing predictions has also been linked to learning in 12-month-old infants in a study that showed superior learning following scenarios that contradicted the laws of object solidity and spatiotemporal continuity (Stahl & Feingenson, 2015). Infants who saw an object go through a wall or appear from a distant location in the scene successfully learnt a new, unrelated property about the object, while infants who witnessed no such violation did not map the same properties onto the object. A follow-up study with pre-schoolers has also shown that children learnt a novel word only when the object had previously defied their knowledge of physical laws by unexpectedly appearing from a new location following a hiding event, suggesting that surprise facilitates learning when the information would otherwise be forgotten (Stahl & Feingenson, 2017).

The effect of surprise on learning might be explained by a heightened attentional state and an increase in arousal levels, which have been shown to enhance memory consolidation and recall in adults (Bradley et al., 1992; McGaugh, 2004). Arousal seems to enhance memory through an increase in noradrenaline levels facilitating the consolidation of the newly acquired information of high salience or emotional significance (McGaugh, 1990; Roozendaal & McGaugh, 2011). Such arousal related increases in noradrenaline are directly correlated with an increase in pupil dilation (Preuschoff, Hart & Einhäuser, 2011; Bradley et al., 2008), offering

a single measure for the cognitive mechanisms underlying both surprise and enhanced encoding.

Furthermore, the violation of a prior prediction might increase the uncertainty in the environment, which has been linked to higher learning rates in adults where participants had to dynamically update their predictions in light of unexpected information (Nassar et al., 2012; McGuire et al., 2014; O'Reilly, 2013). Therefore, the uncertainty resulting from a surprising event might increase the amount of cognitive effort deployed to explore and analyse the unexpected event outcome, which in turn might result in deeper encoding and better memory for the surprising item. In accordance with the idea that pupil size is a reliable measure of cognitive effort (Hess & Polt, 1964; Granholm et al., 1997), increased pupil dilation during encoding has been associated with enhanced memory performance in both adults (Goldinger, He & Papesh, 2009; Papesh, Goldinger & Hout, 2012) and infants (Cheng, Káldy & Blazer, 2019; Káldy & Blazer, 2020).

Changes in arousal levels and/or cognitive effort indicated by increased pupil dilation have been suggested to be mediated by a change in attentional processing systems, with more dilation linked to an increase in focused attention to task-relevant stimuli (Laeng, Sirois & Gredebäck, 2012; Aston-Jones & Cohen, 2005). Time-locked increases in pupil dilation have thus been interpreted as 'interrupt signals' or 'network-reset signals' that indicate a switch in attentional resources when detecting a new target or event (Bouret & Sarah, 2004; Dayan & Yu, 2006). Therefore, if surprise is associated with increased cognitive effort and arousal, which is mediated by an increase in focused attention, pupil dilation following the surprising event might be an index of the depth of encoding and the likelihood of the successful retrieval of these events.

In the current study we investigated the effect of semantic violations on subsequent recognition memory by presenting 17-month-olds with a sequence of familiar images, half of

which were labelled correctly, while the other half were mislabelled. Pupil diameter was measured to establish whether infants responded to the mismatching labels with more dilation compared to the matching labels. Following the naming events, infants saw the previously labelled images along with previously unseen images of familiar objects. Pupil dilation was then used as the measure of recognition memory, with more dilation indicating stronger memory for those objects.

Pupillometry has been used extensively in developmental research to establish a surprise response to unexpected social and physical events (Jackson & Sirois, 2009; Sirois & Jackson, 2012; Gredebäck & Melinder, 2011; Hepach & Westermann, 2016). In the auditory domain increased pupil dilation has been observed in infants and toddlers after the presentation of deviant sounds in an oddball paradigm (Wetzel et al., 2016), in response to animal sounds that did not match the presented image (Krüger, Bartels & Krist, 2019), and following the mispronunciation of familiar words (Tamási et al., 2017; 2019; Fritzsche & Höhle, 2015). Furthermore, pupil dilation has been used as a reliable index of recognition memory in adults (Võ et al., 2008; Otero, Weekes & Hutton, 2011; Goldinger & Papesh, 2012), as well as in infants (Hellmer, Söderlund, & Gredebäck, 2018). Importantly, in the adult literature increased pupil dilation at retrieval has also been observed when the study items were presented acoustically (Otero, Weekes & Hutton, 2011), when the test items were semantically related to the study items (Montefinese, Vinson & Ambrosini, 2018), as well as when the test item was incorrectly judged as old (Kafkas & Montaldi, 2015). This evidence suggests that pupil dilation at retrieval does not simply respond to the perceptual features of the previously presented objects, but rather, it indicates the strength of the underlying memory trace.

Although pupil dilation has been used in infants both to establish the presence of surprise following violations of expectation, as well as an indicator of recognition memory, to our knowledge no study has directly explored the relationship between surprise and learning by

linking a quantifiable index of surprise to the strength of the resulting episodic memory for the object. Importantly, the studies by Stahl & Feingenson (2015; 2017) did not use a direct measure of surprise, but rather inferred its presence or absence from the differences in the dependent variable across conditions.

We predicted an increase in pupil dilation following mislabelling, indicating that infants' lexical expectations were violated. Crucially, if violations of expectation result in a more in-depth encoding of the object, infants are expected to respond with more dilated pupils to the items that were previously mislabelled compared to the previously correctly labelled items at the recognition memory test. On the other hand, if surprise does not enhance infants' memory for the mislabelled items, pupil diameter at test is expected to be different as a function of novelty, but independent of the effect of surprise concerning the previously mislabelled items.

Two different types of auditory violations were used in a between-subjects design that differed in perceptual novelty and the degree to which the violation was impossible or merely unlikely. In the Wrong Label condition familiar objects were mislabelled by a familiar but mismatching name (dog = *banana*), while in the Novel Label condition the familiar objects were mislabelled by an unfamiliar name (dog = *moxie*). While given the relevant lexical knowledge, the former scenario is impossible, the latter scenario may reflect a natural process in language acquisition whereby a previously over-extended category is gradually narrowed down to attain the correct referential scope of the word (Naigles & Gelman, 1995; Gelman et al., 1998). Furthermore, if surprise and subsequent memory enhancement are simply driven by perceptual novelty, the effect is only expected to be present in the Novel Label, but not in the Wrong Label condition.

Methods

Participants

Forty-one typically developing, monolingual 17-month-olds were tested, 23 in the Wrong Label and 18 in the Novel Label condition. (501 – 532 days, $M = 515.63$, $SD = 10.14$, 17 females). Infants were included if they had valid data for at least 50% of the trials in both parts of the experiment. An additional 10 babies were tested but excluded due to insufficient data. Infants were recruited from the Greater London area through the online database of the Centre for Brain and Cognitive Development, Birkbeck. The research received ethical approval from the Ethics Board of Birkbeck, University of London and complied with the guidelines laid down in the Declaration of Helsinki. Written informed consent was obtained from parents or guardians prior to the testing sessions.

Stimuli

Thirty names of familiar objects were selected using Wordbank (Frank et al., 2017). An item was selected if at least 60% of infants understand it at 17 months of age according to the data based on the Oxford CDI (Flocchia, 2017). Thirty phonologically legitimate non-words were generated that were matched to the familiar names in consonant-vowel structure and syllable length using a pseudoword generator (Keuleers & Brysbaert, 2010). Images corresponding to the familiar labels were selected from the stimuli set of two previously published papers on word learning (Bergelson, 2012; 2018) and matched for brightness and saturation using Final Cut Pro. The correct and incorrect labels as well as the corresponding images are displayed in the Supplementary Materials (Table S4). The sequence of the Labelling and Recognition memory trials are depicted in Figure 1.

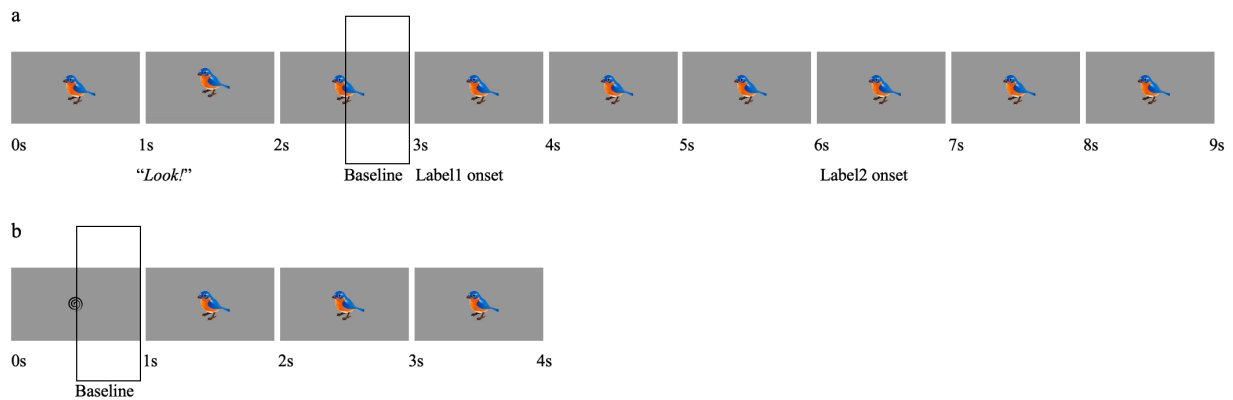


Figure 1. Labelling (a) and Recognition memory test (b) sequences. a) Labelling: The image of a familiar object appeared in the centre of the screen (0s), and it wobbled while the voice said “Look!” (1-2s). Following that the object remained stationary during the two labelling events (3s, 6s). The 500ms interval preceding the onset of the first label was used as a baseline for measuring changes in pupil dilation in response to the labelling events. b) Recognition memory: Following a central fixation stimulus (0-1s) the image appeared in the centre of the screen for 3s (1-4s) with no visual changes and no auditory stimuli during the trial. The 500ms of the fixation stimulus preceding the onset of the test image was used as a baseline to measure changes in pupil dilation in response to the presentation of the object.

Labelling

Each infant was presented with 20 naming events. On each trial a familiar object appeared in the centre of the screen (visual angle: $11^{\circ} 39' 0.45''$) on a grey background (R: 150, G: 150, B:150). The object wobbled slightly (1s-2s) while a pre-recorded female voice said “Look!” in infant directed speech. Following this, the object remained stationary, and the voice labelled the object twice, either with a matching or a non-matching label, depending on the condition.

The two labelling events started at 3s and 6s following the onset of the image, and each utterance lasted approximately 1s. Each label had been recorded twice for each object so that the trial resembles a natural naming event. The two instances of each label always appeared in the same order across babies. Each trial lasted 9 seconds. Trials were separated by a 1s interstimulus interval while a grey screen was presented along with a brief auditory stimulus to orient the infant’s attention to the next trial.

Recognition memory test

Following the labelling events, 30 familiar objects were presented during a silent recognition memory test. Each object was preceded by a central fixation stimulus presented for 1s. The luminosity of each fixation stimulus was matched to the subsequent test image. Following this, the object appeared in the centre of the screen on a grey background and remained stationary for 3s. Each trial lasted 4s. After every 5 trials a short video was played for a maximum duration of 30s to maintain the infant's interest in the trials.

Design

Infants were randomly assigned to either the Wrong Label or the Novel Label group. During the first half of the experiment infants were presented with 20 naming events where a familiar object was labelled. Half of the objects received the matching, familiar label (Bird labelled *Bird*), while the other half of the objects received an unmatching, familiar label (Bird labelled *Car*) or an unmatching, novel label (Bird labelled *Costet*), depending on the group. The order of correct and incorrect labelling, as well as the pairings between objects and incorrect labels were randomised across the trials.

Following the naming events, infants saw a silent recognition memory test where all the objects presented at labelling appeared with 10 new objects. The order of the 30 images was randomised across the memory test. The identity of correctly labelled, incorrectly labelled and novel images were randomised across infants.

Procedure

Infants were tested in a quiet, dimmed room (4.5 lx). Parents were given dark glasses and they were instructed not to interact with their infants throughout the experiment. Stimuli was presented on a Tobii TX300 eye-tracker using Matlab at a sampling rate of 120Hz. The

labelling and the recognition memory test were presented in succession, and participants did not take breaks during the experiment. If the infant was inattentive, auditory and/or visual attention getters were presented for a few seconds during both parts of the experiment. Stimulus presentation was resumed once the infant attended to the screen again. The experiment lasted approximately 10 minutes including calibration.

Data Pre-processing and reduction

Linear interpolation was applied to the pupil data corresponding to each eye with a maximum gap of 10 missing samples (83.33ms). Pupil baseline was established as the 500ms of the stationary image preceding the onset of the first label in the labelling trials, and the 500ms of the fixation image preceding the onset of the test image in the recognition memory trials. Trials were excluded if there was no valid data for either of the eyes for at least 60% of both the baseline and the test intervals after interpolation. Included trials were baseline corrected by subtracting the mean baseline values for each eye from the pupil data of the corresponding eye in the test intervals. Missing data from one eye was replaced with data from the other eye, and the data was subsequently averaged across the two eyes.

The relatively narrow interpolation window and the criterion of 60% valid data within a single trial after interpolation were chosen as a conservative threshold for retaining trials for further analyses. In order to test whether our results are robust after interpolating larger gaps in the data, we ran the same analyses with a 300ms interpolation window, which resulted in the retention of 99 and 35 additional trials during the labelling and the memory intervals, respectively. The conclusions drawn from these results, displayed in the Supplementary Materials, are identical to the findings presented in our main analyses below.

Infants were excluded if they did not contribute at least 10 trials during the labelling phase ($n = 7$) *and* at least 15 trials in the memory phase ($n = 3$). These criteria correspond to,

on average, 5 trials per condition and 50% of the entire stimulus set in each phase. Each participant contributed at least 3 trials to each condition in both parts of the experiment. Infants contributed a total of 712 trials during labelling ($M = 17.36$, $SD = 2.6$, Correct labelling: $M = 8.83$, $SD = 1.56$, Mislabelling: $M = 8.54$, $SD = 1.51$) and a total of 904 trials during the recognition memory test ($M = 22.04$, $SD = 3.85$, Previously correctly labelled: $M = 7.12$, $SD = 2.17$, Previously mislabelled: $M = 6.78$, $SD = 1.75$, New: $M = 8.15$, $SD = 1.49$).

Two intervals of interest were identified: (i) the 6s following the onset of the first label until the end of the trial during the labelling phase (0s: Label1 onset, 3s: Label2 onset), and (ii) the 3s following the onset of the test image in the recognition memory phase.

In order to assess whether pupil dilation differed reliably as a function of condition during these intervals, and to examine the time course of the effects in a data-driven fashion, a permutation analysis was conducted on the pupil data in the two intervals of interest (R, package: *permutes*, $np = 1000$). During the permutation analysis, the condition labels are randomly reassigned to the data numerous times at a given time point, and on each iteration a two-tailed t -test is conducted on this data (or an F -test in case of more than two conditions). The t values over all iterations create a distribution of test statistics values observed under the null hypothesis. The observed t -statistic using the actual condition labels is then compared to the distribution of t values derived from permutation analysis at that time point, and the difference across conditions is deemed significant if the observed t value falls outside the distribution of values that could have occurred by chance (i.e. a p value is significant at 0.05 if less than 50 out of 1000 t -tests had an absolute value larger than the one observed). Permutation testing is repeated at each data point and the resulting test statistics and significance values indicate the time points where the data could not have been obtained if the mapping between the independent and the dependent variable were random. This technique has been used in previous studies analysing pupil dilation in both the infant (Hochmann & Papeo, 2014; Cheng,

Káldy & Blazer, 2019) and the adult literature (Kloosterman et al., 2015; Geng et al., 2015; Quirins et al., 2018).

Since a permutation analysis was performed at each data point throughout the intervals, multiple comparisons were corrected by adjusting the alpha level to an expected false-discovery rate of 5% using the method of Benjamini and Hochberg (1995), a technique that has been widely used to correct for multiple comparisons when analysing pupil data (Einhäuser et al., 2008; Katidioti, Borst & Taatgen, 2014; Lavín, San Martín & Rosales Jubal, 2014; Preuschoff, Hart & Einhäuser, 2011; Mill, O'Connor & Dobbins, 2016). A series of consecutive significant *p* values following the correction indicates a reliable difference across conditions, as well as the time course of the effect throughout the interval of interest.

In addition, based on the results of the permutation analysis we selected a 1s analysis window within the 3s interval following the onset of each labelling event and within the 3s recognition memory phase to conduct inferential statistics. The rationale behind this analysis was to test whether the effects observed as a result of the permutation analysis are robust to changes in interval length, and also to increase comparability with previous infant studies that averaged across a larger time window when analysing pupil dilation in response to violations-of-expectation (Gredebäck & Melinder, 2010; 2011; Krüger, Bartels & Krist, 2019; Pätzold & Liszkowski 2019; Gredebäck et al., 2018) and as an index of recognition memory (Hellmer, Söderlund, & Gredebäck, 2018).

Results

Labelling

The pupil data during the 6s interval following the onset of the first labelling event and the *t* values resulting from the permutation of the data at each timepoint are depicted in Figure 2. The black line (top line) indicates the time points at which the permutation analysis yielded

a significant difference between the pupil values in the two conditions at the initial significance level of $p < 0.05$. The red line (bottom line) indicates the significant time points after correcting for multiple comparisons ($p < p_{FDR=0.05}$). The grey rectangles indicate the intervals where corrected p values remained significant throughout consecutive time points for at least 50ms. The presence of a multiple series of significant p values indicated reliable differences in pupil dilation between the two labelling conditions that were sustained throughout at least 50ms after approximately 1.5s, 2s and 2.5s of the onset of the first label (1533-1591ms, 2008-2091ms, 2716-2791ms) and approximately 2s after the onset of the second label (4950-5033ms).

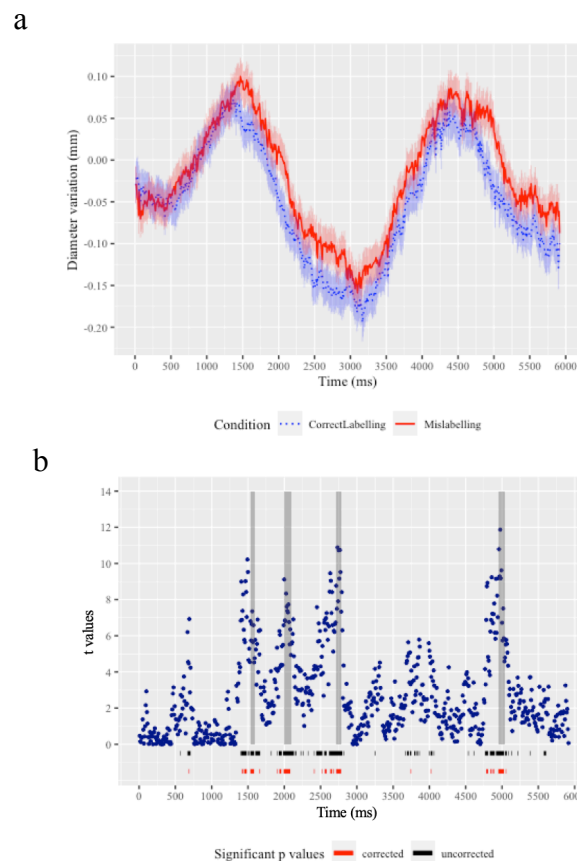


Figure 2. Pupil dilation following correct labelling and mislabelling. a) Changes in baseline corrected pupil diameter as a function of condition. $T=0$ corresponds to the Label1 onset and $t=3$ corresponds to the Label2 onset. Error bars represent the standard error of the mean. b) Results of the permutation analysis indicating significant differences across conditions at each time point (black line/top line: uncorrected, red line/bottom line: following Benjamini & Hochberg correction for multiple comparisons). The shaded areas represent the time points where consecutive corrected p values remained significant for at least 50ms.

Based on the results of the permutation analysis, infants' pupil data in each condition was averaged within the intervals corresponding to the 1.5-2.5s following the onset of each label (1500-2500ms and 4500-5500ms) in order to test whether this effect is reliable throughout a 1s analysis window. The selection of these time windows also allows for the effects of the two labelling events to be tested after the same latencies. This is a more parsimonious approach than postulating a difference between the time course of the two effects.

Average pupil values in the two conditions measured in the two analysis windows were submitted to a 2 x 2 x 2 mixed design ANOVA with the within-subjects variables of Window (1, 2) and Condition (Correct labelling, Mislabelling) and the between subjects variable Group (Wrong Label, Novel Label). The ANOVA yielded a significant main effect of Condition, $F(1,39) = 4.68$, $p = 0.037$, $\eta^2 = 0.107$, with more dilated pupils after mislabelling ($M = 0.007$, $SE = 0.022$) compared to correct labelling ($M = -0.033$, $SE = 0.022$). In addition, there was a significant main effect of Window, $F(1,39) = 4.74$, $p = 0.036$, $\eta^2 = 0.108$, with more dilated pupils in the second window ($M = 0.001$, $SE = 0.023$) compared to the first window ($M = -0.026$, $SE = 0.019$), irrespective of condition. No other effects were significant.

The permutation analysis, as well as the results of inferential statistics confirmed that mislabelling a familiar object with either a mismatching familiar name or an unrelated, novel name elicits increased pupil dilation, suggesting that infants' lexical expectations were violated.

In addition, pupil responses to mislabelling and correct labelling were tested using mixed effects models with Participant and Object as random factors. These models also account for the possibility that infants' familiarity with the objects and the labels may have affected their pupil dilation. In summary, all our findings using mixed models support the hypotheses presented above, and the findings are closely aligned with the results of the ANOVA. These models are reported in the Supplementary Materials.

Recognition memory test

In order to establish whether recognition memory processes could have been influenced by looking time differences during the Labelling phase, a 2 x 2 ANOVA was conducted on infants' total looking time following the onset of the first label during the labelling trials with the within subjects variable Condition (Correct labelling, Mislabelling) and the between subjects variable Group (Wrong Label, Novel Label). The ANOVA yielded a main effect of Group, $F(1,39) = 4.861$, $p = 0.033$, $\eta^2 = 0.111$, with longer looking times in the Wrong Label group ($M = 5327.44$, $SE = 93.17$) compared to the Novel Label group ($M = 5017.39$, $SE = 105.32$), irrespective of Condition. No other effects were significant. Importantly, there was no significant difference between infants' looking time to the images that were labelled correctly compared to the images that were mislabelled, $F(1,39) = 1.99$, $p = 0.166$, $\eta^2 = 0.049$, Correct labelling: $M = 5227.17$, $SE = 81.17$, Mislabelling: $M = 5117.67$, $SE = 79.44$.

In order to investigate whether infants' memory was enhanced for those objects that were previously mislabelled, permutation analysis was performed on the pupil data during the Recognition memory phase using the 3s of the still image following baseline, as described above. Pupil dilation in response to the three conditions during the memory phase and the F statistics with the corresponding significance values at each time point are displayed in Figure 3.



Figure 3. Pupil dilation in response to the three conditions: previously correctly labelled, previously mislabelled, new. a) Changes in baseline corrected pupil values in the three conditions. T=0 corresponds to the image onset. Error bars represent the standard error of the mean. b) Results of the permutation analysis performed at each time point (black line/top line: uncorrected, red line/bottom line: following Benjamini & Hochberg correction for multiple comparisons). The shaded areas represent the time points where consecutive corrected p values remained significant for at least 50ms.

As with the labelling trials, a 1s analysis window was selected to confirm that pupil dilation differed across conditions and to compare infants' responses to the previously correctly labelled, previously mislabelled and new images. Based on the results of the permutation analysis, the pupil data between 2-3s after the onset of the images was submitted to a 3 x 2 ANOVA with the within-subjects variable Condition (Previously correctly labelled, Previously mislabelled, New) and the between-subjects variable Group (Wrong Label, Novel Label).

The ANOVA yielded a significant main effect of Condition, $F(2,78) = 4.23$, $p = 0.018$, $\eta^2 = 0.098$. Bonferroni corrected pairwise comparisons revealed a significant difference between previously correctly labelled and new items ($p = 0.017$) and between previously

mislabeledled and new items ($p = 0.047$), with no difference between the previously correctly labeledled and mislabeledled items ($p = 1.000$). (Correct: $M = 0.013$, $SE = 0.025$, Mislabeledled: $M = 0.002$, $SE = 0.027$, New: $M = -0.075$, $SE = 0.022$). The ANOVA also resulted in a main effect of Group, $F(1,39) = 5.86$, $p = 0.02$, $\eta^2 = 0.131$, with more dilated pupils in the Wrong Label ($M = 0.019$, $SE = 0.021$) compared to the Novel Label group ($M = -0.059$, $SE = 0.024$), irrespective of Condition. No other effects were significant.

Similarly to the pupil dilation data during labelling, we tested infants' responses during the recognition memory phase using mixed effects models with Participant and Object as random factors. These results are displayed in the Supplementary Materials, and the findings are closely aligned with the results of the ANOVA reported above.

In order to directly investigate the differences in recognition memory between the previously mislabeledled and correctly labeledled items, we compared the memory scores in these two conditions with a two-sided Bayes Factor t -test (BF_{10}) using a JZS prior = 0.707 (R, package: *BayesFactor*). The resulting Bayes factor of $BF_{10} = 0.180$ indicates substantial evidence that the means in these conditions did not differ. In other words, the data are $1/BF_{10} = 5.54$ times more likely to have occurred under the null than under the alternative hypothesis.

Furthermore, to test the relationship between pupil dilation after mislabelling and infants' memory for the images, we ran two correlation analyses between the difference scores in pupil size during the labelling trials in each window of interest (Mislabelling – Correct labelling in Window 1 and 2, respectively) and the difference scores at the memory test (Previously mislabeledled items – Previously correctly labeledled items). If surprise enhances memory for the item, pupil dilation during encoding is expected to be positively correlated with pupil dilation during the recognition memory test.

There was no significant association between the responses to the violation and the memory scores when using the labelling data in the first window, $r(39) = -0.150$, $p = 0.348$. In

addition, there was a significant negative correlation between increased pupil dilation following the second label and pupil dilation as an index of memory, $r(39) = -0.310$, $p = 0.049$. However, this latter finding should be interpreted with caution, as the p values are not adjusted to account for the two comparisons. In addition, we conducted the same correlation analyses on the data interpolated with a maximum tolerated gap of 300ms, which are presented in the Supplementary Materials, and we found no association between infants' pupil dilation following correct and incorrect labelling and their memory scores in response to these items.

Overall, the correlations indicate the absence of a positive relationship between the pupil values used to assess the effect of surprise and the pupil values used to assess recognition memory at test.

Discussion

In this experiment we tested whether infants' surprise response elicited by the violation of their lexical knowledge can be linked to their subsequent memory for the objects using pupil dilation as an index of both surprise and recognition memory.

Our main goal was to investigate whether the increased arousal, greater cognitive effort, and ultimately an increase in focused attention elicited by the incorrect labelling events enhanced the processing of the surface features of the mislabelled items. The unexpected label in the mislabelling condition might have also elicited a motivation in infants to reassess their category knowledge and attempt to resolve the conflict between the label and the referent. Although these processes may have occurred, they are not strictly necessary for infants to show better memory for these objects, but an increased attention following mislabelling alone could have resulted in a memory advantage.

Firstly, we replicated the findings of earlier literature suggesting that infants respond to violations of expectation with increased pupil dilation (Jackson & Sirois, 2009; Sirois &

Jackson, 2012; Gredebäck & Melinder, 2011; Tamási et al., 2017; 2019; Fritzsche & Höhle, 2015). These responses measured in pupil diameter might signal an increase in arousal levels, greater cognitive effort, or an increase in focused attention when processing the mislabelled items.

In accordance with previous studies, the presence of a surprise response in the Wrong Label group indicates that pupil dilation is not merely an index of stimulus novelty, but it is also sensitive to violations that do not involve the presentation of unfamiliar stimuli, but rather a mismatch between two perceptually familiar items.

Interestingly, a previous study with 30-month-olds did not find differences in pupil dilation to correct compared to mismatching familiar labels or non-words (Fritzsche & Höhle, 2015). In our study we used infant-directed speech (“*Look!*”) combined with the contingent movement of the object at the beginning of the trial, which might have helped infants form the referential link between the label and the object, whereas in the above study the object remained stationary throughout the trial. However, it is also possible that by 30 months of age toddlers’ lexical expectations are already firmly established, which might prevent them from associating the label with the object, or it might result in the outright rejection of the mismatching label. This explanation is in line with the idea that learners allocate their attention to events that have an intermediate complexity (Kidd et al., 2012; 2014), possibly discarding new information if the discrepancy between the prediction and the outcome exceeds a certain threshold.

Secondly, we replicated previous findings in the adult (Võ et al., 2008; Otero, Weekes & Hutton, 2011; Goldinger & Papesh, 2012) and in the infant literature (Hellmer, Söderlund, & Gredebäck, 2018) showing that participants respond with increased pupil dilation to previously seen compared to unseen items. In this experiment, a single, brief presentation of 20 items elicited reliable differences in infants’ recognition memory between previously seen and unseen items, offering an insight into episodic memory performance at 17 months of age.

Crucially, despite evidence of increased pupil dilation in response to mislabelling, we found no difference in pupil size during the recognition memory test that would indicate that infants had better memory for those items that had previously violated their expectations. This finding is at odds with adult studies using similar paradigms that suggest enhanced memory for items that were encoded under surprising or arousing circumstances (McGaugh, 1990; Vö et al., 2008; Nassar et al., 2012; Roozendaal & McGaugh, 2011).

Importantly, our results do not support the hypothesis previously proposed by Stahl & Feingenson (2015; 2017) that violations of expectation enhance learning about the surprising object. However, there are a few important differences between our design and the paradigms used by Stahl & Feingenson (2015; 2017), which might have contributed to the lack of evidence to support a link between surprise and improved learning. Firstly, in our study the trials that accorded with infants' expectations and those that violated them were presented equiprobably in a random order; therefore, the overall uncertainty regarding a subsequent violation was essentially maximised. Studies with adults suggest that the anticipation of uncertainty, as well as the infrequent violation of a previously generated set of predictions has the most robust effect on learning (Nassar et al., 2012; McGuire et al., 2014). In line with this reasoning, the repetition of surprising events might also impair the resulting encoding benefit. Although we found strong evidence for the presence of a surprise response throughout the experiment, presenting fewer surprise trials may have been more effective in increasing infants' memory performance.

Importantly, the above studies that suggest a link between surprise and learning in infants and pre-schoolers (Stahl & Feingenson, 2015; 2017) used single trials in a between-subjects design that left no room for the participant to accommodate to the uncertainty of the experimental setting. In contrast, our design was chosen in a way to disentangle the effect of the infrequency of the stimuli from the effect of surprise itself. On the one hand, it is possible that infants did not encode the surprising items better due to the repetition of the violations,

while on the other hand they may have shown increased encoding for both types of items due to the relative uncertainty of the environment. Future studies investigating the role of surprise in learning should attempt to disentangle the individual effects of violations of expectation from the underlying change in cognitive mechanisms that occur throughout the experiment.

Although surprise may have enhanced the encoding of both types of items to a certain extent, the large number of the images and the short presentation times make it highly unlikely that infants' overall memory performance was at ceiling. Earlier studies with pre-schoolers using similar set sizes of items in recognition memory tasks did not find performance to be at ceiling even at 3 years of age (Parkin & Streete, 1988; Balcomb & Gerken, 2008). Therefore, the lack of a difference between the previously correctly labelled and mislabelled items at the memory test cannot be explained by a near perfect learning of the entire stimulus set. In addition, pupil dilation is a reliable index of the strength of the underlying memory trace, rather than a binary measure (Otero, Weekes & Hutton, 2011; Montefinese, Vinson & Ambrosini, 2018; Kafkas & Montaldi, 2015), therefore, a memory enhancement for the surprising items could have been observed even if the correctly labelled items were also remembered to some degree.

In summary, although our study offers positive evidence regarding infants' detection of semantic violations, as well as the presence of recognition memory processes, we found no evidence for a positive relationship between infants' surprise response and their memory for these objects. In contrast, we found a negative correlation between pupil dilation after mislabelling and pupil diameter in response to the image during the recognition memory test. However, this association was only observed using one of the analysis windows, and the significance value was not corrected for multiple comparisons. Future studies should investigate whether certain types of surprise may hinder the encoding of the surprising material.

In conclusion, while surprise may facilitate learning and memory in some circumstances, an enhanced encoding of the object for further memory is not a necessary outcome of surprise.

References

- Aslin, R. N. (2007). What's in a look?. *Developmental Science*, *10*(1), 48-53.
- Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annal. Review of Neuroscience*, *28*, 403-450.
- Baillargeon, R. (1986). Representing the existence and the location of hidden objects: Object permanence in 6-and 8-month-old infants. *Cognition*, *23*(1), 21-41.
- Baillargeon, R., Spelke, E. S., & Wasserman, S. (1985). Object permanence in five-month-old infants. *Cognition*, *20*(3), 191-208.
- Balcomb, F. K., & Gerken, L. (2008). Three-year-old children can access their own memory to guide responses on a visual matching task. *Developmental science*, *11*(5), 750-760.
- Begus, K., Gliga, T., & Southgate, V. (2014). Infants learn what they want to learn: Responding to infant pointing leads to superior learning. *PloS One*, *9*(10), e108817.
- 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)*, *57*(1), 289-300.
- Bergelson, E., & Swingle, 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.
- Bergelson, E., & Swingle, D. (2018). Young infants' word comprehension given an unfamiliar talker or altered pronunciations. *Child Development*, *89*(5), 1567-1576.
- Bonawitz, E. B., van Schijndel, T. J., Friel, D., & Schulz, L. (2012). Children balance theories and evidence in exploration, explanation, and learning. *Cognitive Psychology*, *64*(4), 215-234.
- Bouret, S., & Sara, S. J. (2004). Reward expectation, orientation of attention and locus coeruleus medial frontal cortex inter- play during learning. *European Journal of Neuroscience*, *20*, 791–802.
- Bradley, M. M., Greenwald, M. K., Petry, M. C., & Lang, P. J. (1992). Remembering pictures: pleasure and arousal in memory. *Journal of experimental psychology: Learning, Memory, and Cognition*, *18*(2), 379-390.
- Bradley, M. M., Miccoli, L., Escrig, M. A., & Lang, P. J. (2008). The pupil as a measure of emotional arousal and autonomic activation. *Psychophysiology*, *45*(4), 602-607.
- Charlesworth, W. R. (1964). Instigation and maintenance of curiosity behavior as a function of surprise versus novel and familiar stimuli. *Child Development*, 1169-1186.
- Charlesworth, W. R. (1966). Persistence of orienting and attending behavior in infants as a function of stimulus-locus uncertainty. *Child Development*, 473-491.

- Cheng, C., Káldy, Z., & Blaser, E. (2019). Focused attention predicts visual working memory performance in 13-month-old infants: A pupillometric study. *Developmental Cognitive Neuroscience, 36*, 100616.
- Cook, C., Goodman, N. D., & Schulz, L. E. (2011). Where science starts: Spontaneous experiments in preschoolers' exploratory play. *Cognition, 120*(3), 341-349.
- Dayan, P., & Yu, A. J. (2006). Phasic norepinephrine: A neural interrupt signal for unexpected events. *Network, 17*, 335–350.
- Einhäuser, W., Stout, J., Koch, C., & Carter, O. (2008). Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proceedings of the National Academy of Sciences, 105*(5), 1704-1709.
- Floccia, C. (2017). Data collected with the Oxford CDI over a course of 5 years in Plymouth Babylab, UK. With the permission of Plunkett, K. and the Oxford CDI from Hamilton, A., Plunkett, K., & Schafer, G., (2000). Infant vocabulary development assessed with a British Communicative Development Inventory: Lower scores in the UK than the USA. *Journal of Child Language, 27*, 689-705.
- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of child language, 44*(3), 677-694.
- Fritzsche, T., & Höhle, B. (2015). Phonological and lexical mismatch detection in 30-month-olds and adults measured by pupillometry. In The Scottish Consortium for ICPHS (Ed.), *Proceedings of the 18th International Congress of Phonetic Sciences*, Glasgow, Scotland. Retrieved from <http://www.icphs2015.info/pdfs/Papers/ICPHS0339.pdf>
- Gelman, S. A., Croft, W., Fu, P., Clausner, T., & Gottfried, G. (1998). Why is a pomegranate an apple? The role of shape, taxonomic relatedness, and prior lexical knowledge in children's overextensions of apple and dog. *Journal of Child Language, 25*(2), 267-291.
- 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, 9*, 435, 1-14.
- 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.
- Goldinger, S. D., He, Y., & Papesh, M. H. (2009). Deficits in cross-race face learning: insights from eye movements and pupillometry. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(5), 1105-1122.
- Granholm, E., Morris, S. K., Sarkin, A. J., Asarnow, R. F., & Jeste, D. V. (1997). Pupillary responses index overload of working memory resources in schizophrenia. *Journal of Abnormal Psychology, 106*(3), 458-467.
- Gredebäck, G., & Melinder, A. (2010). Infants' understanding of everyday social interactions: A dual process account. *Cognition, 114*(2), 197-206.

- Gredebäck, G., & Melinder, A. (2011). Teleological reasoning in 4-month-old infants: pupil dilations and contextual constraints. *PLoS One*, *6*(10), e26487.
- 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, 1-11.
- Hellmer, K., Söderlund, H., & Gredebäck, G. (2018). The eye of the retriever: developing episodic memory mechanisms in preverbal infants assessed through pupil dilation. *Developmental Science*, *21*(2), e12520.
- Hepach, R., & Westermann, G. (2016). Pupillometry in infancy research. *Journal of Cognition and Development*, *17*(3), 359-377.
- Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611), 1190-1192.
- Hochmann, J. R., & Papeo, L. (2014). The invariance problem in infancy: A pupillometry study. *Psychological science*, *25*(11), 2038-2046.
- Jackson, I., & Sirois, S. (2009). Infant cognition: going full factorial with pupil dilation. *Developmental Science*, *12*(4), 670-679.
- Kafkas, A., & Montaldi, D. (2015). The pupillary response discriminates between subjective and objective familiarity and novelty. *Psychophysiology*, *52*(10), 1305-1316.
- Káldy, Z., & Blaser, E. (2020). Putting Effort Into Infant Cognition. *Current Directions in Psychological Science*, 0963721420903015.
- Katidioti, I., Borst, J. P., & Taatgen, N. A. (2014). What happens when we switch tasks: Pupil dilation in multitasking. *Journal of experimental psychology: applied*, *20*(4), 380, 1-17.
- Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627-633.
- 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.
- Kloosterman, N. A., Meindertsma, T., van Loon, A. M., Lamme, V. A., Bonnef, Y. S., & Donner, T. H. (2015). Pupil size tracks perceptual content and surprise. *European Journal of Neuroscience*, *41*(8), 1068-1078.
- Krüger, M., Bartels, W., & Krist, H. (2019). Illuminating the Dark Ages: Pupil Dilation as a Measure of Expectancy Violation Across the Life Span. *Child Development*. <https://doi.org/10.1111/cdev.13354>

- Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious?. *Perspectives on psychological science*, 7(1), 18-27.
- Lavín, C., San Martín, R., & Rosales Jubal, E. (2014). Pupil dilation signals uncertainty and surprise in a learning gambling task. *Frontiers in Behavioral Neuroscience*, 7, 218,1-8.
- Mather, E., & Plunkett, K. (2011). Same items, different order: Effects of temporal variability on infant categorization. *Cognition*, 119(3), 438-447.
- McGaugh, J. L. (1990). Significance and remembrance: The role of neuromodulatory systems. *Psychological Science*, 1(1), 15-25.
- McGaugh, J. L. (2004). The amygdala modulates the consolidation of memories of emotionally arousing experiences. *Annual Review of Neuroscience*, 27, 1-28.
- McGuire, J. T., Nassar, M. R., Gold, J. I., & Kable, J. W. (2014). Functionally dissociable influences on learning rate in a dynamic environment. *Neuron*, 84(4), 870-881.
- 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.
- Montefinese, M., Vinson, D., & Ambrosini, E. (2018). Recognition memory and featural similarity between concepts: the pupil's point of view. *Biological Psychology*, 135, 159-169.
- Naigles, L. G., & Gelman, S. A. (1995). Overextensions in comprehension and production revisited: Preferential-looking in a study of dog, cat, and cow. *Journal of Child Language*, 22(1), 19-46.
- Nassar, M. R., Rumsey, K. M., Wilson, R. C., Parikh, K., Heasley, 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. (2013). Making predictions in a changing world—inference, uncertainty, and learning. *Frontiers in Neuroscience*, 7(105), 1-10.
- Otero, S. C., Weekes, B. S., & Hutton, S. B. (2011). Pupil size changes during recognition memory. *Psychophysiology*, 48(10), 1346-1353.
- Papesh, M. H., Goldinger, S. D., & Hout, M. C. (2012). Memory strength and specificity revealed by pupillometry. *International Journal of Psychophysiology*, 83(1), 56-64.
- Parkin, A. J., & Streete, S. (1988). Implicit and explicit memory in young children and adults. *British Journal of Psychology*, 79(3), 361-369.
- Pätzold, W., & Liszkowski, U. (2019). Pupillometry reveals communication-induced object expectations in 12-but not 8-month-old infants. *Developmental science*, 22(6), e12832.
- Preusschoff, K., Hart, B. M., & Einhauser, W. (2011). Pupil dilation signals surprise: Evidence for noradrenaline's role in decision making. *Frontiers in Neuroscience*, 5(115) 1-12.

- Quirins, M., Marois, C., Valente, M., Seassau, M., Weiss, N., El Karoui, I., ... & Naccache, L. (2018). Conscious processing of auditory regularities induces a pupil dilation. *Scientific Reports*, 8(1), 1-11.
- Roozendaal, B., & McGaugh, J. L. (2011). Memory modulation. *Behavioral Neuroscience*, 125(6), 797-824.
- Sirois, S., & Jackson, I. R. (2012). Pupil dilation and object permanence in infants. *Infancy*, 17(1), 61-78.
- Stahl, A. E., & Feigenson, L. (2015). Observing the unexpected enhances infants' learning and exploration. *Science*, 348(6230), 91-94.
- Stahl, A. E., & Feigenson, L. (2017). Expectancy violations promote learning in young children. *Cognition*, 163, 1-14.
- Tamási, K., McKean, C., Gafos, A., & Höhle, B. (2019). Children's gradient sensitivity to phonological mismatch: considering the dynamics of looking behavior and pupil dilation. *Journal of Child Language*, 46(1), 1-23.
- 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.
- Twomey, K. E., & Westermann, G. (2018). Curiosity-based learning in infants: a neurocomputational approach. *Developmental Science*, 21(4), e12629.
- Võ, M. L. H., Jacobs, A. M., Kuchinke, L., Hofmann, M., Conrad, M., Schacht, A., & Hutzler, F. (2008). The coupling of emotion and cognition in the eye: Introducing the pupil old/new effect. *Psychophysiology*, 45(1), 130-140.
- Wetzel, N., Buttelmann, D., Schieler, A., & Widmann, A. (2016). Infant and adult pupil dilation in response to unexpected sounds. *Developmental Psychobiology*, 58(3), 382-392.
- Wynn, K. (1992). Addition and subtraction by human infants. *Nature*, 358(6389), 749-750.

Supplementary Materials

I. Analysis of the dataset using an interpolation window of 300ms

In order to test whether the size of the window used for interpolation in the main analysis (83.33ms) contributed to our findings, we ran the same analyses on the dataset after interpolating missing pupil values with a maximum gap of 36 samples (300ms). This resulted in the retention of 1 additional participant, and 99 and 35 additional trials during the labelling and the memory intervals, respectively. This larger sample was then submitted to the same data pre-processing steps and statistical analyses as reported in the Results section.

Labelling data

Pupil data in the labelling phase interpolated with a maximum gap of 36 missing samples (300ms), and the results of permutation analysis are depicted in Supplementary Figure 1. The t values indicate the differences across conditions at each time point. The black lines (top lines) indicate the significant (uncorrected) p values, and the red lines (bottom lines) indicate the time points where differences across conditions remained significant after correcting for multiple comparisons using the method of Benjamini and Hochberg (1995).

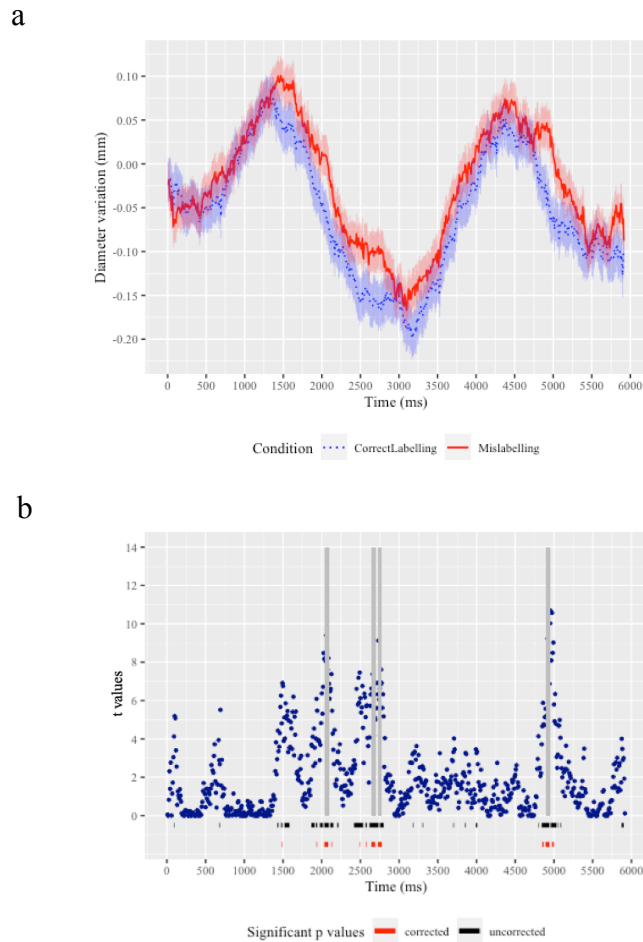


Figure S1. Pupil dilation following correct labelling and mislabelling using an interpolation window of 300ms. a) Changes in pupil dilation as a function of condition. 0s: Label1 onset, 3s: Label2 onset. Pupil data was baseline corrected using the 500ms of the stationary image preceding the onset of the first label. Error bars represent the standard error of the mean. b) Results of the permutation analysis ($n_p = 1000$) indicating significant differences ($p < 0.05$) across conditions at each time point (black line/top line: uncorrected, red line/bottom line: corrected).

Intervals where consecutive corrected p values remained significant for at least 50ms (grey rectangles) indicated a difference between conditions emerging after 2000ms of the onset of the first label (2033-2091ms, 2641-2700ms and 2725-2775ms) and approximately 1800ms following the onset of the second label (4891ms-4950ms). Based on this analysis, we selected two 1s intervals to conduct inferential statistics and assess whether these differences are robust across a larger time window, as discussed in our main analysis.

Infants' pupil data in each condition was averaged in the 2000-3000ms interval following the onset of the first label and in the 1500-2500ms following the onset of the second label. A 2 x 2 x 2 mixed design ANOVA was conducted on the pupil data with the within-

subjects variables of Window (1, 2) and Condition (Correct labelling, Mislabelling) and the between subjects variable Group (Wrong Label, Novel Label). This analysis yielded a significant main effect of Condition, $F(1, 40) = 4.51$, $p = 0.040$, $\eta^2 = 0.101$, with more dilated baseline-corrected pupils following mislabelling ($M = -0.040$, $SE = 0.025$) compared to correct labelling ($M = -0.080$, $SE = 0.023$). The analysis also yielded a significant main effect of Window, $F(1,40) = 51.42$, $p < 0.001$, $\eta^2 = 0.562$, with more dilated pupils in the second window ($M = -0.015$, $SE = 0.025$) compared to the first window ($M = -0.105$, $SE = 0.021$), irrespective of Condition. No other effects were significant.

In addition, we tested whether the effect is also present in the time windows reported in our main analysis, and we conducted the ANOVA using the same independent variables on the pupil data averaged within the 1500-2500ms intervals following the onset of each labelling event (1500-2500ms, 4500-5500ms). This analysis also resulted in a significant main effect of Condition, $F(1,40) = 4.69$, $p = 0.036$, $\eta^2 = 0.105$. Baseline-corrected pupil diameter was significantly larger following mislabelling ($M = 0.001$, $SE = 0.022$) compared to correct labelling ($M = -0.039$, $SE = 0.024$). No other effects were significant.

In conclusion, the interpolation of missing data using the larger window of 36 samples (300ms) and the smaller tolerated gaps of 10 samples (83.33ms) resulted in the same findings, and the effect was also robust in the analysis windows that differed from the ones used in the main analysis.

Recognition memory data

Infants' pupil data during the memory phase interpolated with a maximum gap of 36 missing samples (300ms) and the results of permutation analysis are depicted in Supplementary Figure 2. Intervals where corrected p values remained significant for consecutive time points across at least 50ms (grey rectangles) indicated a difference across conditions emerging

approximately 1600ms after the onset of the test image, which was sustained for a large proportion of the remaining interval (1608-1750ms, 1791-1891ms, 1933-2000ms, 2133-2250ms, 2300-2516ms, 2600-2741ms, 2816-2891ms).



Figure S2. Infants pupil data during the recognition memory test. a) Changes in pupil dilation as a function of condition. 0s: Image onset. Pupil data was baseline corrected using the 500ms of the fixation image preceding the onset of the test image. Error bars represent the standard error of the mean. b) Results of the permutation analysis ($n_p = 1000$) indicating significant differences across conditions at each time point.

As with the labelling data, in order to test whether these differences were robust across a larger interval, a 1s analysis window was selected to conduct further inferential statistics. Based on the results of the permutation analysis, infants' pupil dilation in the three conditions were averaged within the analysis window of 2000-3000ms following the onset of the test image and submitted to a 3 x 2 mixed design ANOVA with the within-subjects variable

Condition (Previously correctly labelled, Previously mislabelled, New) and the between-subjects variable Group (Wrong Label, Novel Label). This analysis yielded a significant main effect of Condition, $F(2,80) = 3.63$, $p = 0.031$, $\eta^2 = 0.083$. Infants pupils were significantly more dilated in response to the previously correctly labelled compared to the new images, $t(41) = 2.55$, $p = 0.015$, and the previously mislabelled compared to the new images, $t(41) = 2.61$, $p = 0.012$, with no difference between the previously correctly labelled and mislabelled images, $t(41) = 0.026$, $p = 0.980$. (Correct: $M = 0.018$, $SE = 0.025$, Mislabelled: $M = 0.017$, $SE = 0.025$, New: $M = -0.056$, $SE = 0.021$). In addition, the ANOVA also resulted in a significant main effect of Group, $F(1,40) = 4.33$, $p = 0.044$, $\eta^2 = 0.098$, with more dilated pupils in the Wrong Label group ($M = 0.023$, $SE = 0.022$) compared to the Novel Label group ($M = -0.44$, $SE = 0.024$), irrespective of Condition. No other effects were significant.

In summary, the analysis of the recognition memory data interpolated with a maximum gap of 36 samples (300ms) indicated a similar time course of the effect and resulted in the same statistical findings as our main analysis.

Lastly, we conducted correlation analyses between the difference scores in pupil size during the labelling trials in each window of interest (Mislabelling – Correct labelling in Window 1 and 2, respectively) and the difference scores at the memory test (Previously mislabelled items – Previously correctly labelled items), as in the main analysis. With regards to the first window, we used the Labelling data averaged within both the 1500-2500ms and the 2000-3000ms intervals following the onset of the first label. If surprise enhances memory for the item, pupil dilation during encoding is expected to be positively correlated with pupil dilation during the recognition memory test. We found no significant associations between infants' pupil dilation following correct and incorrect labelling and their memory scores in response to these images: Window1 (1500-2500ms): $r(40) = -0.058$, $p = 0.717$, Window1 (2000-300ms) : $r(40) = -0.101$, $p = 0.525$, Window2 (4500-5500ms): $r(40) = -0.254$, $p = 0.105$.

These results are in line with our conclusions presented in the main analysis, namely that infants did not show enhanced recognition memory for those images that had previously been mislabelled.

II. Analysis using mixed effects models

In order to test whether individual differences or differences in the response to the objects influenced our results, we first built a simple linear model with fixed effects only, and then added 1) Participant with a random intercept 2) Object with a random intercept 3) allowing the effect of Condition to vary across participants (Participant with random slope) 4) allowing the effect of Condition to vary both across participants and across objects (Participant with random slope and Object with random slope). (R, packages: *lmer*, *lmerTest*.)

At each step we assessed model performance by computing model deviance with χ^2 statistics to test whether the addition of the random part to the model improved overall performance. We also tested the predictors for significance to test whether adding the random effects to the model diminished the explanatory power of the fixed effects. Our findings regarding the Labelling and the Recognition memory data are displayed in Table S1 and Table S2, respectively.

Labelling data

For the Labelling data, we first built a simple linear model with the fixed effects of Condition, Window, Experiment, and their interactions. This resulted in a significant model: $F(7, 1429) = 2.37$, $p = 0.02$, with Condition as a significant predictor: $F(1, 1429) = 8.92$, $p = 0.0028$. No other predictors approached significance. Therefore, the interaction terms were dropped from the subsequent models, and we used the simple model with the fixed effects of

Condition, Window and Experiment as a starting point for adding the random parts to the subsequent models.

The χ^2 tests show that model performance improved with the addition of each random effect, and pupil responses varied across participants and across objects. However, after allowing the intercepts and slopes of Participant and Object to vary, Condition still emerged as a significant predictor in all of the models (with the exception of the last model where the effect of Condition was marginally significant, $p = 0.052$).

Therefore, although there was a variability in participants' responses to the violations, as well as a variability in the effect depending on the objects that were mislabelled, this variability alone cannot account for the main effect of Condition observed in our analysis in the paper.

Table S1. Labelling data

Fixed effects	Random effect(s)	Log-likelihood	Model deviance compared to previous model	Predictor: Condition (no other predictors approached significance)
Condition Group Window	-	-235.53		$F(1,1433) = 8.92$ $p = 0.0028$
Condition Group Window	Participant: random intercept	-164.13	$\chi^2(1) = 142.78$ $p < 0.00001$	$F(1,1397.01) = 10.82$ $p = 0.0011$
Condition Group Window	Participant: random intercept Object: random intercept	-147.11	$\chi^2(1) = 34.05$ $p < 0.00001$	$F(1,1389.7) = 11.95$ $p = 0.0006$
Condition Group Window	Participant: random intercept + slope Object: random intercept	-142.74	$\chi^2(2) = 8.73$ $p = 0.012$	$F(1,40.5) = 6.27$ $p = 0.016$
Condition Group Window	Participant: random intercept + slope Object: random intercept + slope	-139.17	$\chi^2(2) = 7.14$ $p = 0.028$	$F(1,32.74) = 4.06$ $p = 0.052$

Recognition memory data

For the Memory data, we first built a simple model with the fixed effects of Condition, Group and their interaction. This yielded a significant model: $F(3, 1129) = 5.06$, $p = 0.001$, with Condition and Group as significant predictors. The interaction term was not significant.

(Condition: $F(1, 1129) = 8.63, p = 0.003$, Group: $F(1, 1129) = 6.03, p = 0.014$.) The ANOVA reported in the paper also yielded a main effect of Group, with more dilated pupils in the in the Wrong Label group compared to the Novel Label group, irrespective of Condition.

Similarly to our approach with the Labelling data, the interaction term was dropped from the subsequent models, and we added the random effects to the simple model with the fixed effects of Condition and Group. The same random effects were used as with the Labelling data, and the model deviance statistics and the effect of the significant predictors are displayed in Table S2 below.

Adding Participant and Object with random intercepts both improved the model, but adding random slopes did not result in any further improvement. The largest improvement was observed when Object was added with a random intercept, which strengthens the assumption that pupil dilation is a reliable measure of recognition memory, and the objects that the babies had been familiar with prior to the experiment may have elicited larger dilation than the ones they had encountered for the first time during the Labelling phase.

Importantly, Condition emerged as a significant predictor in all of the models, indicating that prior familiarity with the objects alone cannot account for the variability in pupil size during the recognition memory test.

Table S2. Memory data

Fixed effects	Random effect(s)	Log-likelihood	Model deviance compared to previous model	Predictor: Condition	Predictor: Group
Condition Group	-	-318.05		$F(1,1130) = 8.64$ $p = 0.0033$	$F(1,1130) = 6.03$ $p = 0.014$
Condition Group	Participant: random intercept	-314.86	$\chi^2(1) = 6.44$ $p = 0.011$	$F(1,1100) = 8.23$ $p = 0.0042$	$F(1,42.9) = 3.91$ $p = 0.054$
Condition Group	Participant: random intercept Object: random intercept	-288.49	$\chi^2(1) = 52.7$ $p < 0.00001$	$F(1,1082.8) = 8.33$ $p = 0.0039$	$F(1,42.4) = 3.59$ $p = 0.064$
Condition Group	Participant: random intercept + slope Object: random intercept	-287.64	$\chi^2(2) = 1.69$ $p = 0.42$	$F(1,41.1) = 7.01$ $p = 0.011$	$F(1,42.2) = 4.05$ $p = 0.051$
Condition Group	Participant: random intercept + slope Object: random intercept + slope	-286.22	$\chi^2(2) = 2.85$ $p = 0.24$	$F(1,38.2) = 6.19$ $p = 0.017$	$F(1,42.3) = 4.09$ $p = 0.049$

We also tested whether adding participant and object as random factors would change our conclusions regarding the two conditions of primary interest, namely if differences would emerge between the previously mislabelled and previously correctly labelled items during the memory test. To this end, we followed the same modelling steps on the Memory data described above *excluding* the condition that involved new items. These results are displayed in Table S3 below.

We first built a simple model using Condition (Previously mislabelled/Previously correctly labelled), Group and their interaction as fixed effects. This did not result in a significant model overall: $F(3,715) = 1.99$, $p = 0.114$. However – similarly to the model above and in the ANOVA reported in the manuscript – the fixed effect of Group was significant: $F(1,715) = 5.02$, $p = 0.025$. No other predictors approached significance.


Importantly, the fixed effect of Condition (excluding responses to new items) did not approach significance either in the simple model, or in any of the subsequent models after the









addition of the random parts to the models. This strengthens our conclusions reflected in the ANOVA and the Bayesian statistics reported in the manuscript, namely that there was no difference between infants' memory performance as a function of the type of labelling they had previously heard. Instead, in the Memory data, familiarity alone accounts for the effect of Condition reported in the ANOVA in the main analysis and in the mixed models above (i.e. previously seen items elicited larger pupil dilation compared to new items, irrespective of whether they had been labelled correctly or incorrectly).

Table S3. Memory data excluding the condition with new items

Fixed effects	Random effect(s)	Log-likelihood	Model deviance compared to previous model	Predictor: Condition	Predictor: Group
Condition (Mislabelled/correct) Group	-	-171.12		$F(1,716) = 0.17$ $p = 0.67$	$F(1,716) = 5.03$ $p = 0.025$
Condition (Mislabelled/correct) Group	Participant: random intercept	-170.23	$\chi^2(1) = 1.79$ $p = 0.187$	$F(1,686.7) = 0.19$ $p = 0.655$	$F(1,42.4) = 4.10$ $p = 0.049$
Condition (Mislabelled/correct) Group	Participant: random intercept Object: random intercept	-161.43	$\chi^2(1) = 17.6$ $p < 0.001$	$F(1,676.6) = 0.31$ $p = 0.573$	$F(1,41.4) = 3.91$ $p = 0.054$
Condition (Mislabelled/correct) Group	Participant: random intercept + slope Object: random intercept	-160.27	$\chi^2(2) = 2.32$ $p = 0.31$	$F(1,42.4) = 0.21$ $p = 0.655$	$F(1,41.5) = 3.67$ $p = 0.062$
Condition (Mislabelled/correct) Group	Participant: random intercept + slope Object: random intercept + slope	-160.25	$\chi^2(2) = 0.03$ $p = 0.984$	$F(1,42.2) = 0.20$ $p = 0.656$	$F(1,41.5) = 3.68$ $p = 0.061$

Table S4. Correct and incorrect labels and the corresponding images used during Labelling. Incorrect labels were randomly assigned to images in the mislabelling trials.

Image	Word	Non-word
	<p><i>apple</i></p>	<p><i>alster</i></p>
	<p><i>ball</i></p>	<p><i>gull</i></p>
	<p><i>balloon</i></p>	<p><i>cattoun</i></p>
	<p><i>banana</i></p>	<p><i>samana</i></p>
	<p><i>bath</i></p>	<p><i>kir</i></p>
	<p><i>bed</i></p>	<p><i>phim</i></p>
	<p><i>bird</i></p>	<p><i>pirk</i></p>

	<p><i>book</i></p>	<p><i>nook</i></p>
	<p><i>bottle</i></p>	<p><i>cumble</i></p>
	<p><i>bubbles</i></p>	<p><i>bembles</i></p>
	<p><i>car</i></p>	<p><i>daaf</i></p>
	<p><i>cat</i></p>	<p><i>wug</i></p>
	<p><i>chair</i></p>	<p><i>mair</i></p>
	<p><i>cheese</i></p>	<p><i>sheen</i></p>
	<p><i>coat</i></p>	<p><i>sooph</i></p>

	<i>biscuit</i>	<i>costet</i>
	<i>cup</i>	<i>dax</i>
	<i>dog</i>	<i>rog</i>
	<i>door</i>	<i>beel</i>
	<i>duck</i>	<i>dodds</i>
	<i>hat</i>	<i>vak</i>
	<i>milk</i>	<i>dalk</i>
	<i>monkey</i>	<i>bongie</i>

	<p><i>nappy</i></p>	<p><i>moxie</i></p>
	<p><i>phone</i></p>	<p><i>rhove</i></p>
	<p><i>shoe</i></p>	<p><i>bew</i></p>
	<p><i>sock</i></p>	<p><i>zav</i></p>
	<p><i>spoon</i></p>	<p><i>slain</i></p>
	<p><i>teddy bear</i></p>	<p><i>tomalair</i></p>
	<p><i>toothbrush</i></p>	<p><i>timbrook</i></p>

