Pupil Dilation as a Substrate for Cognitive Load in Monolingual and Bilingual Infants: A Pupillometry Study on the Bilingual Advantage

Berg, Karina Elise van den

Pupil Dilation as a Substrate for Cognitive Load in Monolingual and Bilingual Infants: A Pupillometry Study on the Bilingual Advantage

Karina Elise van den Berg

Under the supervision of Dr Andreea Geambaşu & Dr Leticia Pablos Robles

Second Reader: Professor Claartje C. Levelt

Research Master Linguistics

Leiden University

July 2023

Universiteit
Leiden
The Netherlands

Acknowledgements

My gratitude goes towards all the young participants, their parents and their caregivers for making this endeavour possible. Thank you to Dr Sybren Spit for helping out with organizing the data files and to Thomas Tienkamp for providing resources for the analysis.

Special thanks to my friends, particularly Elise Alberts, Natasja Delbar, Sophia Nauta and Aranka van Tol, for all the advice, emotional support and all over good times in the past years. I am glad to have you in my life and to have shared the simple joys of cups of tea and good conversations. I also want to thank my fellow students in my thesis writing group for their support, understanding, and words of encouragement along the way. You made me feel less alone when in trouble. To my parents, thank you for your continued support, even when I felt adrift in the ocean.

Most importantly, I want to express my gratitude to my supervisors, Dr Andreea Geambaşu & Dr Leticia Pablos Robles, and all who have provided aid and supported me along the long road to finalizing this thesis. Your never-ending patience and words of encouragement mean more to me than I can express in words. Thank you for believing I could complete this project. I would not have seen the end of this without your help.

Abstract

The existence of a bilingual advantage in cognitive processing is a popular research topic and is heavily debated. A seminal study by Kovács and Mehler (2009) provided evidence that there is a bilingual advantage in 7-month-old pre-verbal infants in a switching task, although replication findings are inconsistent (Dal Ben et al., 2022; D'Souza et al., 2020; Kalashnikova et al., 2021; Spit et al., 2023). Measuring the pupil dilation response (PDR), a physiological measurement linked to the locus coeruleus (LC) and as such, to cognitive processing load, could give us a more direct look into whether 7-month-old bilingual infants indeed have a cognitive advantage compared to monolinguals.

For this pilot study, additional pupil size measurements were taken in the Leiden arm of the replication effort of Spit et al. (2023) to examine whether bilinguals have a smaller PDR from baseline. After hearing a syllable pattern (AAB or ABB) the infants had to predict on which side a visual reward appeared. The reward side would be the same for the first nine trials in the pre-switch block. The next nine trials had the other syllable pattern and the reward on the other side in the post-switch block. Finally, the last 18 trials mixed both syllable patterns, retaining their associated reward side. The results suggest no difference in cognitive load between the monolingual and bilingual groups when they needed to relearn to predict the target reward side in the post-switch block, nor was there a difference in mean PDR in the association block. This is in line with the results found in Spit et al. (2023) where anticipatory looking behaviour was examined.

However, an exploratory analysis suggested there was a significantly larger PDR in monolinguals during stimulus presentation in the pre-switch block compared to the post-switch block, indicating monolinguals had a higher processing load in the first block of the experiment. The implications of this are unclear, but might be explained by an effect related to the unfamiliarity of the task stimuli seen in monolinguals only due to different attentional strategies between the groups. Future research should be done with larger sample sizes and more sophisticated statistical modelling.

Keywords: Bilingualism, bilingual advantage, cognitive processing load, pupillometry, infants, inhibitory control

Table of Contents

Introduction

Kovács and Mehler (2009) produced a landmark study in the field of bilingualism: they showed that bilingual infants were able to repress an old rule in favour of learning a new reward pattern compared to monolinguals in an eye-tracking study. Monolingual and bilingual 7-month-old infants were familiarised with a syllabic AAB or ABB pattern (experiment 2 in their paper) with a reward appearing on only one side (either right or left) of the screen in the first half of the experiment. During this first half, both mono- and bilingual infants learned to expect the reward on this side. However, when both the pattern and the reward side were switched, bilinguals were better than monolinguals in learning to inhibit looking at the previously correct side of the screen in favour of the new target side. The authors concluded that this is evidence for enhanced cognitive control in bilinguals, even before they actively speak their two languages themselves, a heavily debated finding also observed in older bilingual populations (Paap et al., 2015).

However, infant behaviour during experiments can be difficult to interpret due to fussiness and the absence of verbal communication. Therefore, adding a non-invasive, temporally sensitive, physiological measurement robustly correlated with cognitive processing load, namely pupil dilation, may give valuable insight into whether the enhanced cognitive control is indeed present in the target population. After all, if behaviour (such as an infant looking at a side of the screen) reflects this finding, then so must a reflex controlled by the infant's autonomic nervous system (such as pupil size changes). Previous pupillometry studies in infants as young as 4 months old have proven to be fruitful and a valuable addition to often equivocal results (Gredebäck & Melinder, 2011; I. Jackson & Sirois, 2009).

The Leiden University Centre of Linguistics (LUCL) participated in a replication of the eye-tracking study of Kovács and Mehler (2009) with an additional association block added after the original pre-switch and post-switch blocks. In this association block, the two syllable patterns and their respective reward sides associated with them were mixed (Spit et al., 2023). The current paper describes an addition to this replication study: a different dependent variable, pupil dilation, is used as an involuntary physiological measurement robustly related to cognitive processing load (Gredebäck & Melinder, 2011; I. Jackson & Sirois, 2009) in order to find out if bilinguals indeed find the switch easier to make than monolinguals, as reflected by smaller pupil size increases from baseline in the bilingual group compared to the monolingual control group. As such, the question for this experiment is whether less cognitive processing effort is needed for bilinguals to inhibit an old response to a pattern in favour of a new one compared to a monolingual control group. In other words, is there a smaller pupil dilation response (PDR) relative to monolinguals in the second and third block, i.e. the post-switch block and the association block, where the infants must inhibit previously learned information in order to select the correct reward side?

Bilinguals are expected to have a smaller PDR compared to their per-trial baseline pupil size than monolingual infants, as pupil size reflects cognitive processing effort. This effect is expected in the post-switch block, where the previous rule pattern must be inhibited to correctly predict the new reward side before the reward shows up. It is therefore expected that the bilingual group has, on average, a smaller pupil size change from baseline in the anticipation phase of the post-switch trials.

Background Literature

The Bilingual Cognitive Advantage in Infants

One of the most popular topics within the field of bilingual research is the debate about the ‘bilingual advantage’ (Paap et al., 2015). The proposed cognitive advantage in bilinguals is related to domain-general cognition instead of just being related to language. The term is generally used for two distinct potential benefits related to being bilingual: it has been stated that multilingualism may have a protective effect on the brain, thereby possibly delaying the onset of mild cognitive impairment and dementia (Anderson et al., 2020; Bialystok et al., 2007). Another potential benefit would be advantageous earlier in life, where it has been claimed that multilingualism improves non-linguistic (i.e. domain-general) cognitive processing[1] (Bialystok, 2008, 2009), such as inhibitory control and attention, which is an executive function used for overriding an otherwise habitual or dominant response to a stimulus (Hilchey & Klein, 2011). Many studies have reported some advantage in bilinguals of various age groups being faster on tasks where conflicting information could interfere with choosing the correct response (Adesope et al., 2010; Bialystok et al., 2005; Donnelly et al., 2015; Sabourin & Vinerte, 2015). It is widely theorised that this advantage in inhibition or attentional control is due to bilinguals' continuous practice in language selection: when their multiple language systems are always active in their mind to some degree, additional effort has to be exerted selecting the appropriate language for the context bilinguals find themselves in (Green & Abutalebi, 2013). However, claims of a cognitive bilingual advantage are heavily debated, due to e.g. publication bias for non-null results in favour of the hypothesis (de Bruin et al., 2014; Paap et al., 2015; Paap & Greenberg, 2013).

Yet some form of a bilingual advantage is found in various studies with various paradigms (Adesope et al., 2010), including a seminal paper by Kovács & Mehler (2009) that found a bilingual advantage in pre-verbal 7-month-old infants. This study provided some evidence

[1] Cognitive processes are processes of information in the mind related to acquiring a knowledge or understanding of one's experience of the environment through thoughts and the senses. Examples are attention, decision-making, memory, perception, reasoning and problem-solving, among numerous others. Language perception and production are also two (categories of) cognitive processes, though these are inherently linguistic in nature.

against the idea that repressing one language in favour of producing another one leads to better inhibitory control in bilinguals (Green, 1998). Green's inhibitory control model suggests that bilinguals always have all their languages active to a certain degree and must thus inhibit the language not relevant at a moment in favour of the one that is being used. For example, a Dutch-English bilingual would have to inhibit their Dutch when talking to an English speaker. It is hypothesised that this continued practice of language inhibition is the cause of bilinguals' improved inhibitory control in non-linguistic settings as well. However, the reasoning behind this relates to bilinguals having to manage languages when *actively producing* one, something infants cannot do yet. Kovács and Mehler's (2009) study with pre-verbal participants thus provided evidence against this idea.

In this pilot study, a group of bilinguals and a monolingual control group participated in three similar experiments: in experiment 1, the 7-month-old infants' attention was directed to an eye-tracker's screen with colourful arrows pointing at the centre, during which they heard trisyllabic pseudowords (consisting of the syllables *le, zo, ri, mo, ni,* and *ve*), followed by an anticipation period of one second. During the anticipation period, two white squares on the left and right side of the screen appeared. After the anticipation period, a toy-like reward showed up in either the left or the right square (counterbalanced across participants). This reward side was maintained for 9 trials, after which the next 9 trials (the post-switch block) showed the reward in the white square on the other side. Therefore, the infants could learn to predict which side to look at for the reward, yet also had to learn to inhibit their previous behaviour after the first nine trials. Both monolinguals and bilinguals were found to learn to predict the reward side in the first half of the experiment, i.e. in the pre-switch block, at the same rate. However, only bilinguals learned to anticipate the opposite reward side during the post-switch trials.

This finding was also found in experiment 2, where the separate reward sides of the pre- and post-switch block were linked to two different syllabic patterns, ABB (e.g. *le-mo-mo*) and AAB (e.g. *ni-ni-ve*) and in experiment 3, where the auditory stimuli of experiment 2 were changed into visual stimuli, such as circles and squares appearing on the screen. Kovács and Mehler (2009) concluded that bilingual infants were already better at response inhibition and cognitive control than monolinguals.

However, other explanations than a greater capability of inhibiting old, irrelevant information are offered. For example, Bialystok and Craik (2022) hypothesise that the difference in executive functioning between bilinguals and monolinguals relates to attentional control, which "serves to maintain current goals in an active state, to facilitate cognitive operations that accomplish these goals, to suppress interference, and to switch processing resources to a different set of operations when it is cognitively beneficial to do so" (Bialystok & Craik, 2022, p. 1252). This is because there is also some evidence that bilinguals perform better than monolinguals not just on tasks requiring response inhibition, but on facilitation tasks as well, among others[2] (Bialystok & Craik, 2022). This cannot simply be due to increased inhibitory control, as that should only increase performance on incongruent trials due to the need to repress the distractor cues (e.g. the peripheral instead of central arrows in the Flanker task). However, these effects could be explained by a broader concept such as attentional control, which also includes inhibition.

Other authors also propose possible mechanisms related to differences in attention between bilinguals and monolinguals: D'Souza et al. (2020) conducted four different experiments

[2] a couple of different other tasks bilinguals have performed better on than monolinguals are working memory tasks, disengagement of attention tasks, and false belief tasks (see Bialystok & Craik, 2022, Table 2).

comparing monolingual and bilingual infants, including a conceptual replication of Kovács and Mehler (2009), in which the ABB and AAB patterns consisted of three geometric shapes being presented sequentially in the middle of the screen. They could not replicate the original findings, as both bilinguals and monolinguals increased the number of correct post-switch predictions across the nine trials after the reward side switch. The three experiments they conducted following the replication provided evidence for bilinguals using a more exploratory attention strategy, wherein they switch attention more frequently and disengage attention easier, presumably using this strategy for seeking new information. The authors state that this possible explanation cannot be confirmed through their experiments. The underlying reason for the differences in attention strategies between the two groups remains unclear.

Regardless of the exact cognitive mechanisms behind the bilingual advantage in non-linguistic executive functioning, both Kovács and Mehler (2009) and the replication by Dal Ben et al. (2022) concluded that this advantage is found in 7-month-old babies. The study conducted by Dal Ben et al. (2022) was not an exact replication, as it used no auditory cues. Instead, the attention grabber in the centre of the screen was a circle fluctuating in diameter, after which the anticipation phase and reward presentation followed. As mentioned earlier, this is in line with Kovács and Mehler's original experiment, as they did not include a testing phase in which both syllable patterns (and their respective reward sides) were randomly presented in the same block. However, this does mean that Dal Ben et al.'s replication is not based on any linguistic stimuli, nor does it leave an opportunity for examining whether the infants are able to distinguish the two presented syllable patterns and connect them to a reward side. Additionally, the authors reach different conclusions depending on the method of analysis: using ANOVA, as was also done in the original experiment, led to results in favour of the bilingual advantage that were weak at best. However, the authors also performed a logistic mixed effects analysis in which effects can be tested on a per-trial basis, instead of arbitrarily binning the different trials into three blocks containing three trials each. This logistic regression analysis presented more statistically robust results in favour of the bilingual advantage. Notably, however, bilingual infants did not learn to look at the reward side as quickly as the monolingual group in the first, pre-switch phase, though by trial 9 this between-group difference was gone. The authors conclude that bilingual infants “build more open and less rigid initial representations of the world, which in turn are easier to update when circumstances change. On the other hand, monolinguals seem to be faster in building and strengthening initial representations, making it harder to update them when circumstances change” (p. 26). This discrepancy in the speed of building different representations may also be yet another explanation of the bilingual advantage as measured in these studies: instead of inhibitory function as the main driver of the effect, simply having a less rigid rule to predict the next reward side means bilinguals can adapt quicker when that rule proves itself redundant and needs to be updated.

The same finding, where bilinguals were slower in the pre-switch phase of the study with learning to predict the correct reward side, but corrected their predictions faster in the post-switch phase of the experiment than monolinguals, was also found when Dal Ben et al. (2022) re-analysed the data sets from the other two studies (D’Souza et al., 2020; Kalashnikova et al., 2021). It must be noted that Dal Ben et al.'s logistic regression re-analysis of those data sets with inattentive participants filtered out led to statistically significant results, contrary to the analysis conducted by the original authors. More precisely, D’Souza et al. (2020) and Kalashnikova et al. (2020)'s visual conditions (with ABB and AAB patterns represented on the screen in geometric shapes) had statistically significant results in Dal Ben et al.’s (2022) re-analysis: monolinguals were better than bilinguals at learning to predict the correct reward side in the pre-switch trials, whereas bilinguals outperformed monolinguals when they had to update their rule and learn to predict the other reward side in the post-switch trials of these visual experiments. Yet a re-analysis of Kalashnikova's (2020) auditory condition (with syllable stimuli similar to Kovács and Mehler's experiment 2, as replicated in this paper) resulted in the opposite outcomes: bilinguals outperformed monolinguals in reward prediction in the pre-switch phase of the experiment, but monolinguals outperformed bilinguals in the post-switch phase. This also contradicts the findings of Kovács and Mehler (2009).

Results of this research paradigm remain inconsistent: recently, Spit et al. (2023) conducted a replication study with four different labs in the Netherlands, using the exact materials as used in experiment 2 of the original study. They analysed the data of 98 infants. In addition to the replication of the pre-switch and post-switch blocks, they also included an association block to explore whether either group of infants was able to predict the reward side based on the previous connection between one syllable pattern and its associated reward side, as this was not included in the original paradigm. This association block contained both syllable patterns of the previous two blocks mixed together, which remained linked to their respective reward side for each individual participant. Using an association block like this makes it possible to examine whether the infants connect the syllable patterns to their respective reward sides in order to correctly anticipate where the reward would show up; in the previous iterations of this paradigm, it was possible to replicate the concept of the study without presenting any meaningful stimuli by simply showing the attention grabber followed by the reward sides (see also Dal Ben et al., 2022). The data in Spit et al. (2023) was extensively analysed through various means: an ANOVA was conducted, as was conducted in the original study, yet more modern techniques were also applied: a linear mixed regression model was used to examine the data more granularly, on a per-trial basis, and statistical significance was tested using the Bayesian approach. Their results indicated that both monolinguals and bilinguals were able to update their predictions in the post-switch block, as the relative number of correct looks before the reward phase increased in this block. However, none of the analyses showed a difference in performance between the monolinguals and the bilinguals. Furthermore, neither group was able to predict the correct reward side based on its associated pattern: infants mostly looked at the correct reward side *after* the reward had already appeared on the screen. As such, (conceptual) replications of the same experiment, but with different stimuli modalities or different methods of analysis, lead to different results, while some of the non-null results need a different explanation than originally proposed.

The above leads to the question of what kind of cognitive process might lead to the bilingual advantage in the Kovács & Mehler replications, if it is found at all. It could be related to inhibitory control, attentional control in general, a more exploratory strategy in attention direction, the strength of initial rule representations or flexibility herein, something else related to task-switching, memory, or other forms or combinations of cognitive processes in the infant's mind. It remains unclear what the bilingual advantage in infants in this setting is supposed to be. However, it is not unlikely that a supposed bilingual advantage in executive function is related to *some* sort of decreased processing effort, as this would be in line with the neural efficiency hypothesis: people with increased cognitive abilities appear to have lower brain activation and experience lower effort on the same tasks than people with lower cognitive abilities than them (Di Domenico et al., 2015; Verney et al., 2004). Moreover, intrasubject differences in cognitive load are found between easy and difficult tasks (Dunst et al., 2014), including in pupillometry research (Hess & Polt, 1964; Kahneman & Beatty, 1966).

Furthermore, the heterogeneity in the execution of the experiments, reliance on different methods of analysis leading to different outcomes, and these different rationalisations of the underlying cognitive mechanisms are not the only reasons why it is difficult to come to consistent conclusions within this same experimental paradigm: infant research is limited to experiment designs that account for both the young participants’ attention span and the inability to communicate the task to the child. A physiological response as an indicator of cognitive load may not solve some of these issues in infant research, but it could provide an additional way to enlighten us on what is happening inside the infant's mind and help with the interpretation of infant research.

Pupillometry

These issues bring us to an explanation of pupillometry research in the field of cognitive psychology. While it is well-known that the pupil responds to changes in its exposure to light, constricting when in a bright environment (i.e. the pupillary light reflex), the psychological relevance of pupil dilation responses (PDRs) in an experimental setting became apparent in the early 1960s, when the first seminal studies on PDRs related to cognition were published (Laeng et al., 2012). Participants that were shown difficult multiplication problems, for example, had more widely dilated pupils than when they were shown easy multiplication problems (Hess & Polt, 1964). Throughout the decades, more studies related pupil dilation to conditions requiring increased cognitive load, such as during tasks related to memory load, decision-making, attention, emotional or cognitive arousal, or situations to do with surprise, conflict, or uncertainty (Joshi & Gold, 2020; Sirois & Brisson, 2014).

This pupil dilation response appears to be present not only in humans: it has also been found in monkeys and rats (Joshi & Gold, 2020). Additionally, it appears consistently across all age groups, from studies in adults (Hershaw & Ettenhofer, 2018; Hess & Polt, 1964; Kahneman & Beatty, 1966) to infants (Hepach & Westermann, 2016; Jackson & Sirois, 2022; Jackson & Sirois, 2009; Ross-Sheehy & Eschman, 2019; Zhang & Emberson, 2020) as young as four months old (Addyman et al., 2014; Gredebäck & Melinder, 2011). For example, a study looking at prediction in adults and 6-month-old infants examined whether both groups showed signs of top-down prediction (i.e. through top-down neural signals without the need for input[3], as opposed to bottom-up prediction guided by sensory inputs) by having them participate in the same omission experiment (Zhang et al., 2019): trials consisted of the presentation of an auditory stimulus, after which a figure was presented on the screen together with another sound. This was followed by a waiting period with a blank screen in order to have a distinct time window for PDRs. There were also omission trials interspersed in between the visual presentation trials. Zhang et al. (2019) found that both adults and 6-month-old infants had greater PDRs when the visual stimulus was omitted. This suggests that infants already have the capacity to make top-down predictions, just like adults.

Jackson and Sirois (2009) studied identity violation of expectation in 8.5-month-old infants by analysing both looking times and pupil size changes. Their experiment consisted of a familiarization phase with 6 trials, in which a video of self-propelling toy trains went around a circular track: the train would disappear into a tunnel, come outside again, ride around across the track, go through the tunnel again to then come to a halt after it emerged from the tunnel a second time. The familiarization trains were red and green. In three testing trials, the following changes took place: a different colour train (blue) went around the track in the same manner (novel yet possible), a familiar train colour went around the track but changed to the other familiar colour once it emerges from the tunnel the second time and comes to a halt (familiar impossible), and one trial shows a train going around the track in a familiar colour at first, to then emerge in the unfamiliar colour (novel impossible). Infants had longer looking times in the possible trials when the trains had a familiar colour, but longer at impossible events when the trains had a new colour. Looking times alone were considered ambiguous due to test-order effects. However, analysis of the pupil dilation response across the length of

[3] As is done by the omission of a stimulus that had been repeatedly presented to the participant before.

the trial showed a distinct interaction effect between novelty and possibility: a violation of expectation was found, shown through pupil size increases, when the train emerged from the tunnel in a different colour, but only when it emerged in the novel colour. This study shows that pupillometry in infant cognition research may be a valuable addition to behavioural (gaze direction, looking time) measurements in eye-tracking studies (Jackson & Sirois, 2009).

The reason why the pupils on their own could provide information on the demands of various cognitive processes is due to their indirect connection to many networks of the brain. Pupil size changes related to arousal and cognition have been linked to the locus coeruleus-norepinephrine (LCNE) neuromodulatory system (Joshi & Gold, 2020; Laeng et al., 2012; Laeng & Alnaes, 2019). The locus coeruleus (LC) is located in the pons of the brainstem and produces norepinephrine (noradrenaline) (Laeng & Alnaes, 2019). It is an essential nucleus in the management of attention, stress, cognitive control, and decision-making, among many other functions. It has connections to many parts of the brain, including areas of the cerebral cortex, such as the dorsolateral and dorsomedial prefrontal cortices, with other (weaker) connections to other cortical areas (e.g. the parietal and temporal cortices). The anterior cingulate cortex (ACC) also seems to be linked to the LC (Joshi & Gold, 2020). As such, the LC receives signals from many parts of the brain and plays an essential role in the allocation of cognitive resources. It influences pupil dilation through its role in the autonomous nervous system (ANS), where the sympathetic nervous system (related to a more active, energised and aroused state, colloquially referred to as the “fight or flight” or “feed and breed” system) and the parasympathetic nervous system (relating to a more relaxed state, the “rest and digest” system), control pupil dilation and contraction, respectively.

As such, the pupils provide a useful window into many task-related cognitive processes in a non-invasive and relatively inexpensive way. Instead of having to set up an intricated EEG or fNIRS installation and attach this to a child with the hope that it does not interfere with their state of being and influence their behaviour, an eye-tracker can take pupillometric measurements, which may already be a part of the experimental paradigm used, as is the case with the current study.

In the Leiden arm of the Spit et al. (2023) replication study of Kovács and Mehler (2009), additional measurements of the pupil sizes of both eyes (in arbitrary units) were added in order to do precisely this: while infant behaviour may be prone to interpretation difficulties, a physiological measurement as a substrate of cognitive processing effort could potentially be a fruitful way of examining whether bilingual infants indeed have a cognitive advantage over monolinguals in the pre-verbal developmental stage. This is done by looking at pupil size changes from individuals throughout the trials of the experiment and comparing them to the baseline pupil size at the start of each trial (Mathôt et al., 2018). It is expected, in accordance with the original study and the theory of the bilingual cognitive advantage, that bilinguals are better at inhibitory control (or task-switching, or visual attention direction) in the post-switch trials of the experiment than monolinguals. This should be reflected in their relatively lower pupil size changes from baseline compared to monolinguals, although a post-switch mean pupil size increase is still to be expected compared to the pre-switch block in both groups due to having to inhibit the previous task-related information. It is hypothesised that the association trials, where the two patterns and their respective reward sides are mixed together, will elicit a greater PDR in both groups due to the increased task difficulty: infants have to recognise the pattern and (in case they have learned the rule) remember which side of the screen the reward will be displayed during the anticipatory time window before the reward shows up. Only the replication study by Spit et al. (2023), which the current study is based on, has examined behavioural outcomes in an association block. They could not find any significant results between the groups. Therefore, it cannot be hypothesised whether some sort of bilingual advantage would be present in this context, although this possibility can be explored.

Methods

The data collection for this study has been conducted during Leiden University's contribution to the multi-centre replication project of Kovács and Mehler (2009), and as such follows the same in-lab methods and procedures as described in Spit et al. (2023), with the exception of having a percentual language cut-off point for bilingual infants. Bilingual classification is more lenient in the current study (see ‘Participants’ section). An addition to the PyGaze (Dalmaijer, Mathôt, & van der Stigchel, 2014) script used in the Leiden lab ensured pupil size data for both eyes were collected.

Participants

Participants consisted of a subset of the infants tested in the multi-lab replication project (Spit et al., 2023) of the original Kovács and Mehler (2009) study, who were all tested at Leiden University. Nineteen infants participated in the experiment, all of which are between the age of 7 months, 1 day and 7 months, 30 days old. The infants were divided into a monolingual (N = 8; 2 were later excluded, see ‘Analysis’ section for the cut-off points for missing data) and a bilingual group (N = 11; 2 were later excluded, see ‘Analysis’). The bilinguals have different language backgrounds, though mostly with Dutch as one of their languages, whereas the other group consists of monolingual Dutch infants. Considering most infants in the Netherlands grow up in an environment where Dutch is the dominant language, most bilingual infants have unbalanced linguistic input. They have less exposure time to their second language (i.e. non-dominant language) compared to their Dutch input, yet will still be classified as bilingual in this study.

Monolingual infants were exposed to their first language at least 95% of the time they were awake[4] (Spit et al., 2023). Bilingualism in the infant group in the original Kovács and Mehler study did not have a defined relative amount of exposure to any of the child's languages. Bilingualism was therefore based on exposure to multiple languages from birth onwards, without specific cut-offs. The current study follows the same principle: as long as an infant is exposed to a second language enough to not be classified as monolingual (see above) as determined by via the language background questionnaire (LBQ), they are considered bilingual. Multilinguals who are exposed to more than two languages are thus also classified as bilinguals. This led to a mean relative L1 exposure of 66.5% (SD = 8.3%) with a range of 55 to 76.3% (see also the 'Pre-Processing' section below). Participant recruitment ran from April 2021 to December 2021 until the Dutch replication project ended (Spit et al., 2023). to COVID restrictions, fewer participants than initially planned (25 per group) for the Leiden lab were recruited. Participants who were born pre-term, defined as a gestation period shorter than 37 weeks, or who were reported to have a visual impairment, who had a history of more than 3 ear infections, or who had an ear infection at the time of the experiment were excluded from data analysis[5]. Participants were recruited through (social) media channels, e-mail, the Babylab Leiden website, a letter sent to recent parents in the municipality of Leiden, daycares, and through the distribution of flyers in the university's region and nearby municipalities.

[4] In practice, all monolingual participants were 100% exposed to Dutch alone according to the language background questionnaires filled in by their parents. Realistically, it is to be expected that some infants had exposure to other languages (e.g. English) through their parents' media consumption (e.g. music, television).

[5] The participants in the Leiden cohort all met these criteria, so none of the participants had to be excluded for these reasons.

Materials

Materials for the experiment were acquired from the authors of the original experiment (Kovács & Mehler, 2009, experiment 2) and the same Language Background Questionnaire (LBQ) is used (the LBQ can be found in the replication project's Open Science Foundation website: https://osf.io/p4dwu/). Both the LBQ and an information letter were provided via e-mail and during the visit to the Babylab. The visiting parent/guardian was asked to read and sign the consent form during their visit and fill in the LBQ in case they did not fill it in at home.

The LBQ consists of general questions about the infant and their family, such as their day of birth in order to calculate their age in days, their gender, the average time spent sleeping per day, weeks of gestation, relevant medical questions regarding ear tubes, ear infections, visual impairments, questions screening for a family history of language and speech disorders, including dyslexia and dysgraphia, and parents' education level. Other questions collect data on the infant's linguistic input, such as the time spent with the child in each of the parents' languages in percentages, and time spent outside the home environment, including relative language input of languages used there.

For the experiment, a Tobii-T160 eye-tracker with a 24-inch screen is used to acquire eye movements and gaze, fixation, and pupil dilation measurements. The calibration and experiment were scripted (Spit et al., 2023)[6] and run through the Python-based (version 2.7.3) PyGaze (version 0.6.0) Open-Source eye-tracking software (Dalmaijer, Mathôt, & van der Stigchel, 2014).

[6] The lines in the script for recording pupil size measurement was added by the author of this work.

Design

The experiment is preceded by a separate script for calibration, in which wiggling toy ducks were shown with an accompanying ‘ringing’ noise. This stimulus showed up in the upper left, lower left, upper right and lower right corners and in the centre of the screen. The toy duck was visible until the researcher pressed a button on the PC’s keyboard. Calibration rounds were done until sufficiently successful through visual inspection of the locations of the gaze fixations[7], after which calibration was accepted and the experiment was started.

The first half of the experiment design was the same as experiment 2 as described by Kovács and Mehler (2009). Additionally, to test whether the infants learned to associate one reward side with the corresponding stimulus pattern, an additional association phase of 18 randomised trials was also added to the experiment (Spit et al., 2023). Thus, the experiment consisted of three blocks with a total of 36 trials: 9 trials in block 1, 9 trials in block 2 and 18 trials in block 3. The experiment consisted of switch tasks, where a tri-syllabic pseudo-word with either an AAB or ABB pattern was paired with a reward that shows up on one side of the screen (e.g. the AAB pattern is always paired with the reward appearing on the left side of the screen, and the ABB pattern with the reward on the right side). The reward consisted of three different pictures of toy puppets, resembling a star, a bug-like creature or a hippo, paired with an attention-grabbing noise which was played twice in a row, namely a ‘tring’ sound when the star or hippo appeared and a ‘beep-beep’ sound for the bug. The rewards were not associated

[7] The gaze fixations should spatially approximate the position of their respective calibration points: no more than one single calibration point in a corner of the screen was allowed to have a visually detectable skew away from its respective calibration picture (yet still subjectively rated to be ‘close enough’ in proximity by the researcher). The centre calibration point should always have a precise gaze fixation, as the attention grabber in the experiment will only disappear when looked at directly (see below).

with a specific syllable pattern or reward side. The reward appeared in either the left or right white square, which were present on the screen for the entirety of the experiment. Between each trial, the infant's attention was redirected to the centre of the screen with a beeping sound and four colourful arrows rapidly appearing clockwise one after the other (red, blue, green, yellow), pointing at the centre, arranged together like a fixation cross. The formation of this 'cross' by the arrows had a duration of 1 second, after which they also disappeared clockwise in 1 second. This repeated itself until the arrows attract the infant's attention for at least a continuous 500 ms. Afterwards, the arrows disappeared and the tri-syllabic pseudoword with an AAB or ABB pattern and a length of 1.7 seconds is played. These tri-syllabic pseudowords consisted of syllables without a coda, namely combinations of *le, zo, ni* (A syllables) and *mo, ri, ve* (B syllables), all with a duration of 400 ms and with a pause of 250 ms in between the syllables. After the auditory presentation of the pattern, a 1 second anticipation period began in which only the two white boxes on either side of the screen were visible. The corresponding reward appeared at the end of the anticipation period for 2 seconds. The toy puppet rewards switched back and forth between being bigger and smaller each 500 ms to make the reward more interesting.

The first two blocks consist of nine trials in which one pattern (e.g. AAB) is repeatedly shown with the corresponding reward appearing on the same side. The second block consists of the other pattern (e.g. ABB) with the reward only showing up on the other side of the screen. In these blocks, the stimulus pattern and reward presentation sides are counterbalanced across participants such that participants with an even participant number had the reward on the left side of the screen in the first block, whereas participants with an odd participant number had the reward on the right in the first block. In the second half of the experiment (the third block), the 18 trials contain a randomised order of both syllabic patterns still linked to their respective reward sides: in this association phase, the side of the reward is not predictable without paying attention to the pattern presented. This association phase could therefore test to see if infants can connect a syllable pattern with a reward side (see Spit et al., 2023 for the behavioural, i.e. gaze fixation, results) instead of only learning to look at a reward side during a block. This association block was not done in previous iterations of this experimental paradigm. Measurements collected by the eye-tracker are collected in 8.3 ms (at 120 Hz) intervals during the entirety of the experiment.

Procedure

Caregivers of participants were contacted and informed by e-mail after recruitment before making an appointment for the experiment. They received a letter with general information about the procedure and the Language Background Questionnaire (LBQ) to fill in, in addition to a COVID-19 symptom questionnaire. They were instructed to bring a face mask due to the lab's COVID-19 regulations. All infants were tested using an eye-tracker in the baby lab at the Faculty of Social Sciences at Leiden University, the Netherlands. After arrival, the caregiver and infant were led to the lab by the researcher where they could take a seat and where they were allowed to remove the face mask. The infant either remained on the caregiver’s lap with a toy or, if preferred by the caregiver, was placed on a play mat. The researcher explained the procedure of the experiment to the caregiver, informed them about their right to cease participation at any time, and offered the information letter once more to read. Any remaining questions were answered by the researcher and the caregiver was requested to sign the consent forms for their infant’s participation in the experiment and collection and storage of their anonymised data. One consent form is filed by the researchers and the other is for the caregiver’s administration. Contact details were included on the consent form to give caregivers the option to revoke their consent after their visit.

After informing the caregiver and signing the consent forms, the caregiver was requested to turn personal devices to silent mode, after which they were led to the experimental cubicle with the eye-tracker, where they took a seat with their infant on their lap. Due to COVID-19 restrictions, the cubicle's curtain remained open for ventilation and the caregivers did not receive masking headphones or sunglasses. Instead, the experimenter asked them to avoid looking at the eye-tracker's screen to avoid having the eye-tracker follow their eyes instead of their infant's eyes. They could choose to do this by turning their face up or to the side, or to close their eyes. They were instructed to comfortably hold their infant and interact as little as possible with their infant for the entirety of the experiment, unless the infant was fussy or otherwise distracted. In that case, the caregiver could gently correct the position of their infant's torso by recentring them to a neutral position on the caregiver's lap if the infant had turned around. They were not allowed to redirect the infant's head itself. Infants sat at approximately 60 cm away from the eye-tracker, room lights were dimmed to their lowest setting and window curtains were closed, after which the experimenter started the calibration and experiment, respectively. In case of too much fussiness[8] (e.g. crying or too much moving around), the researcher would cease the experiment at the current trial and data up to that

[8] The experimenter or caregiver (whoever determines to cease the experiment first) could decide whether the current participant was not able to continue with the experiment. This decision was not automatically determined by the script after a certain time, though the experiment could not continue if the infant was not paying attention to the experiment due to the attention grabber requiring a minimum of 500 ms of fixation before moving on to the next trial. If an infant was inattentive or crying, the caregiver could soothe the child if necessary. The caregiver was also allowed to recentre the infant's torso towards the screen if they had turned around towards the caregiver. The caregiver was not allowed to influence the infant's head position/gaze. In the case these efforts failed, the researcher would manually abort the experiment.

point would still be saved (see ‘Pre-Processing’ section for the 30% missing data cut-off point used to exclude trials and participants). After the experiment ended, the caregiver and infant were led back to the table in the welcome area. If the LBQ was not filled in and acquired before the experiment before participants came into the lab, the caregiver was given a hard copy of the LBQ to fill in after the experiment. This kept the period between arrival and the experiment as brief as possible and reduce distractions for the infant. Participants were rewarded by being gifted a children’s book, a ‘baby diploma’, and a reimbursement for the travel costs if necessary.

Pre-Processing

A continuous measurement such as pupil size data, which also has absent data points for when the eye-tracker could not measure pupil size, must be transformed before analysis. Pre-processing of the dataset before statistical analysis is necessary for removing measurement artefacts that may negatively affect the analysis. These artefacts are e.g. instances of blinks, missing data points, and other outliers in pupil size highly unlikely to be related to actual pupil size changes, but rather caused by measurement errors (Mathôt et al., 2018; Mathôt & Vilotijević, 2022).

Pre-processing of the data before statistical analysis was done with the help of the *GazeR* package 0.1 (Geller et al., 2020) in R 4.2.3 (R Team, 2014). This package provides pre-processing steps for pupil-related datasets and follows widely regarded recommendations in the field for cleaning up the data (Jackson & Sirois, 2009). After merging the individual data files of all participants[9], a column with the average pupil size was created by averaging the

[9] Special thanks to Dr. Sybren Spit. Of course, all mistakes are my own.

pupil size measurements of both eyes and taking missing values into account. Missing samples of pupil size data of one eye are usually either interpolated from the data point of the opposite eye (Jackson & Sirois, 2009) or interpolated from the average of the previous three and following three data points of that same eye (Jackson & Sirois, 2009). For our analysis, the first option was chosen. After this, the average pupil size of both eyes was calculated for each sample at all time intervals. The participant background data (e.g. language background, age) was separated from the eye-tracking data in order to process the trial data on its own. Rows in the dataset measured outside the time window of the experiment (-500 to 4700 ms) were removed to discard data points irrelevant to the trial itself. Afterwards, trials and participants with more than 30% missing data were removed, which excluded four participants (two bilinguals, two monolinguals[10]), leaving the bilingual group at N=9 (4 girls, m age = 228.3 days, SD = 6.7, range 215 – 241 days; mean relative L1 exposure = 66.5%, SD = 8.3%, range = 55 – 76.3%) and the monolingual group at N=6 (3 girls, mean age = 230 days, SD = 5.7, range 221 – 236 days), with 435 trials spread across 15 participants. Additional removal of trials containing more than 640 sampling points or less than 600 sampling points were also removed, as this indicates eye-tracker-related sampling issues that could interfere with modelling: most trials had around 625 sampled time points. This led to 16 additional trials being removed. Thus, 419 individual trials across 15 participants were kept after these filtering steps. Samples for the time axis were properly aligned for all trials and participants by interpolating the timepoint data onto a common axis. This is done to make sure all trials are comparable on the time axis.

[10] One of the monolinguals (participant 17 in the dataset) was excluded from the analysis due to an unknown error.

Blinks detected by *GazeR* were removed and pupil data was extended over the gaps with fill-back and fill-forward set at 100 ms. The pupil size data is further smoothed out before interpolation. Various options are mentioned in the literature for doing this, such as a low-pass filter with e.g. a sample-to-cut frequency ratio of 10 (Winn et al., 2018) to 12.5 (I. Jackson & Sirois, 2009; Kret & Sjak-Shie, 2019), a Hanning filter, an n-point (e.g. 5-point in Geller et al., 2020) moving average filter, or a median filter (Forbes, 2020). One of the options given by the *GazeR* package was used for this, namely a 5-point average window.

Afterwards, missing data between two pupil measurements were interpolated. There are two commonly used options for interpolation: linear interpolation and cubic-spline interpolation. Mathôt et al. (2018) recommend cubic-spline interpolation using four equally spaced points around the blink. However, Jackson and Sirois (2009) use linear interpolation, while Geller et al. (2020) state differences between the two methods are negligible. During the pre-processing of our dataset, linear interpolation is used, as cubic-spline interpolation led to some extreme values at the onset and offset of the trials. Finally, after these steps, the baseline pupil size of each trial is calculated. In order to compare changes in pupil size to a baseline, it is generally recommended to use baseline pupil size subtraction instead of division, which is done on a per-trial basis (Mathôt et al., 2018). For this data set, the average pupil size during the last 250 ms that the attention grabber is on the screen was calculated. This baseline pupil size is then subtracted from the pupil size during the rest of the trial to calculate pupil dilation compared to this baseline.

Analysis

The collected data was analysed in R version 4.2.3 (Team, 2014) with the *lme4* (Bates, Mächler, et al., 2015), *permutes* (Voeten, 2022), *permuco* (De Rosario-Martinez, 2022) and *phia* (Frossard & Renaud, 2021) packages for linear mixed models (LMMs), cluster permutation testing and post-hoc analysis, respectively, among other supporting packages. A relatively simple mixed model was chosen for the analysis of this data. Although others have argued for using a maximal model (Barr et al., 2013), such a comprehensive model with all possible random slopes and intercepts would contain an unnecessarily large number of fixed and random effects to individually test for statistical significance, in addition to also leading to convergence issues (Bates, Kliegl, et al., 2015; Matuschek et al., 2017).

As many potential random slopes and intercepts are not considered theoretically relevant for our analysis, these are left out when testing the random effects models. To elaborate, one can argue that the study design allows for five main fixed effects, of which three are truly theoretically significant. These are group (bilinguals vs. monolinguals), block (pre-switch, post-switch, and association blocks), trial phase (onset, stimulus presentation, anticipation period, reward), pattern condition (AAB vs. ABB syllable patterns), and reward side (left vs. right side of the screen). The first two main effects, group and block, are important main effects in previous research with a similar study design (D'Souza et al., 2020; Kalashnikova et al., 2021; Kovács, 2009). The third main effect, trial phase, is hypothesised to influence pupil size throughout the trial, where it is expected that cognitive processing within participants starts to increase during the presentation of the stimulus due to having to process the auditory stimulus and continues to increase during the anticipation phase because of infants having to predict where the reward will show up. This means they have to recall previous instances of the reward location and, in the case of trials after the pre-switch block, have to inhibit the previously learned pattern and redirect attention to the new side. This is hypothesised to increase cognitive processing effort. During these phases, pupil size should dilate compared to the attention grabber phase.

Pattern condition might potentially lead to different pupil size changes if infants are particularly attuned to this, although the exact syllable pattern presented should not matter for the outcome, as this experiment could also be performed with visual patterns (D’Souza et al., 2020; Kovács & Mehler, 2009) or even without pattern presentation (Dal Ben et al., 2022). However, previous studies within this paradigm have never shown indications of this effect. Considering the non-unanimous results of these studies on whether bilinguals have better inhibition in the first place, the assumption is that subtle effects of pattern differences are unlikely to be significant. Finally, reward side can be included to control for potential environmental factors, as the opening to the booth and the rest of the room was always on the infant’s right side. If participants show a particular bias to this side, this may have been the reason. However, these expectations are not based on any theoretical foundation or empirical results from similar studies with infants.

Thus, the linear mixed model that was created to test the statistical significance of independent variables and interactions thereof contained the following factors relevant to the hypothesis: the participants’ monolingual or bilingual status (group), and the block that the trial belonged to, i.e. whether a trial was part of the pre-switch, post-switch or the association trials (mixed) block. The phase of the trial (stimulus presentation phase, anticipation phase, or reward phase) was not included as an independent variable in the models, as creating models for the entirety of the trial consistently led to models with non-normally distributed residuals or convergence issues. Instead, models were created with the data from each separate phase of the trial. As such, the trial duration is binned per each trial phase. The first phase, starting when the infant is focused on the centre of the screen, contains the attention grabber and consists of the first 500 ms (i.e. before the onset of the auditory stimulus and is used for baseline calculation). This phase is not analysed as it is used to calculate baseline pupil size per trial. The second phase, which contains the presentation of the syllable pattern, starts at 0 ms and consists of the following 1700 ms of the trial. The third phase is the anticipation period, with two empty squares on the screen for reward prediction. This is the primary phase of interest and is 1000 ms long. Finally, the fourth phase is the reward phase, where the puppet and accompanying jingle are presented, which is 2000 ms long. The entire trial thus is 4700 ms in length. The entirety of the trial consisting of all four phases is the same duration as in the original Kovács and Mehler (2009) experiment, as the first half of the experiment was an exact replication of their work. Kovács and Mehler (2009) also used the same phase distinction (see Kovács & Mehler, 2009, Figure 1), but only the anticipatory phase was used for collecting data on looking behaviour for their analysis.

In addition to the fixed main and interaction effects, random effects were added for subjects and items (‘syllable pattern’). Random intercepts for both subject and item were added to account for participant differences and items leading to different effects. In addition, a random slope for subject over block was added, as it is plausible that different participants have different (relative) reactions to the three blocks. No random slope for subject over group was added as these variables are highly dependent on each other (all subjects belong to one group). Other random effects were left out for the sake of simplicity: adding many random effects, such as for reward side, would lead to a complex model unable to properly converge (Bates, Kliegl, et al., 2015; Matuschek et al., 2017). Therefore, only the hypothetically relevant and plausible random effects discussed above were included in the model. The dependent variable of interest was the mean pupil size of both eyes at a particular point in the trial compared to its baseline at the start of the trial. Infants who do not meet the requirements (see section ‘Participants’) were excluded from the analysis.

There are various manners in which pupillometry data between participant groups can be analysed. In recent literature, more advanced methods such as generalised additive mixed models and functional data analysis are used and recommended, among others (Jackson & Sirois, 2009; van Rij et al., 2019). However, for the analysis, I will use the relatively more straightforward linear mixed-effects model analysis, as using advanced statistical methods is beyond the scope of this thesis. Linear mixed models provide enough statistical rigour to determine whether the independent variables have a significant effect on the PDR and are also frequently used in pupillometry studies (see e.g. de Vries et al., 2023; Gingras et al., 2015; van den Berg et al., 2022). LMMs are also often used for analysing other experimental time series data with fluctuations across trials, such as is also done in EEG studies (Bosma & Pablos Robles, 2020; Heise et al., 2022). Furthermore, cluster permutation tests are performed for the individual models for each phase of the trial in order to find the specific regions of interest—the specific time windows in the phase—where the independent variables lead to statistically significant results (Voeten, 2021).

The difference between cognitive processing effort in mono- and bilinguals in during the trials is indirectly measured by the relative differences between the baseline-corrected mean pupil sizes of the participants in both groups.

Following from the above, the linear mixed model used for significance testing for each phase of the trials is as follows: *mean_pupil ~ Group*Block + (1+Block|subject) + (1|SyllablePattern).* Note that in this model, mean pupil size for statistical modelling is calculated through grouping by participant group and block, instead of all individual data points for pupil size per subject and trial. The latter was used initially, but led to modelling issues where assumptions were not met, such as non-normalised residuals, as modelling in this way tried to fit too many data points.

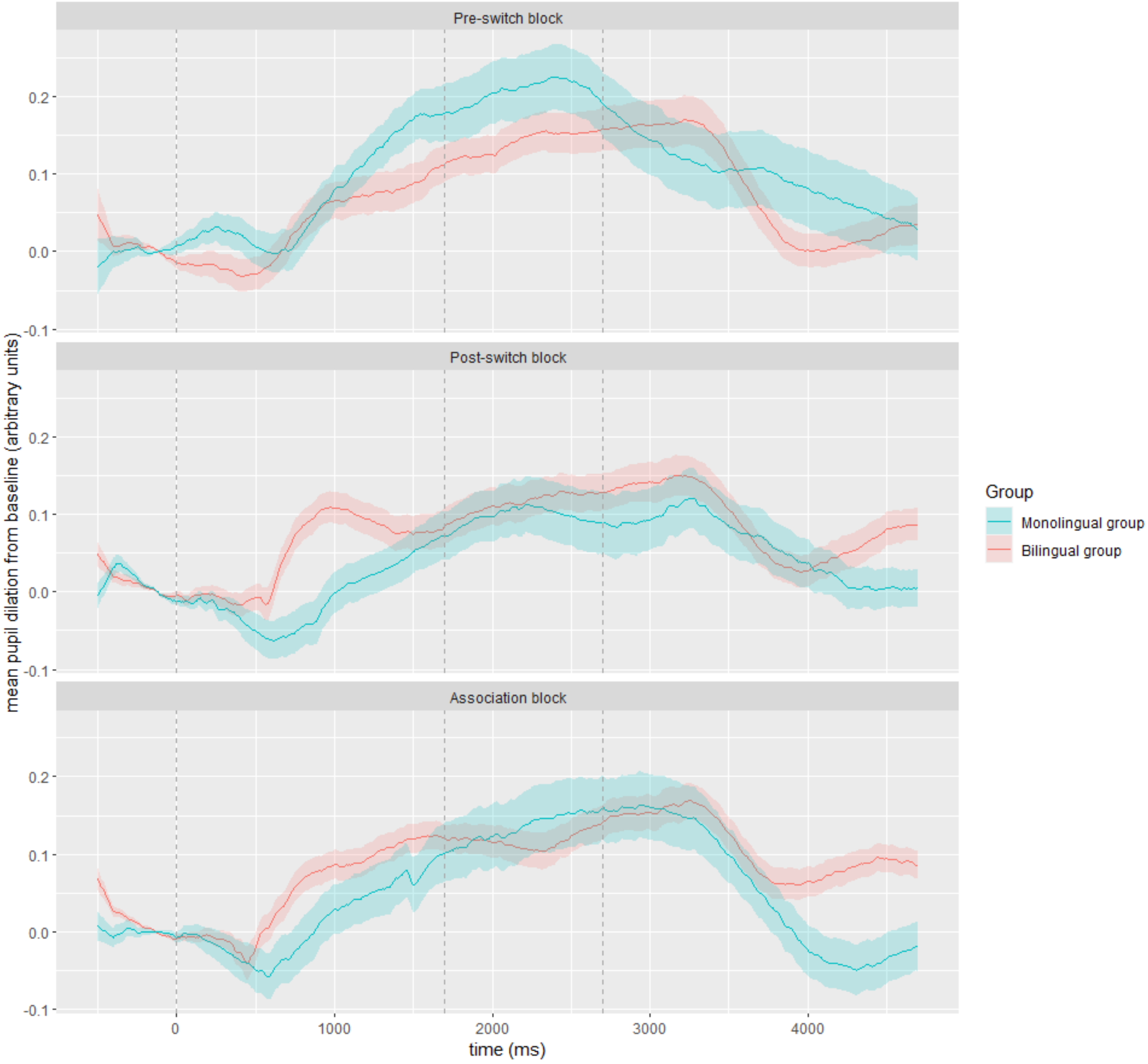
While a Bayesian approach to significance testing was used in the analysis of the multi-centre behavioural data (Spit et al., 2023), this approach is considered beyond the scope of this thesis and thus, frequentist significance testing with the commonly used cut-off value of $p < 0.05$ is used to determine the significance of individual and interaction effects in the model.

Results

As described above, linear mixed models were used to test the significance of the effects of theoretical interest for each individual phase of the trials. The hypothesis states that the main phase of interest is the anticipation phase, in which participants were expected to predict where the reward would show up based on the previously presented auditory stimulus. The stimulus phase and the reward phase will be analysed in the same manner as the anticipation phase, although it should be noted that the analyses for these phases are exploratory in nature as no hypotheses were established for these phases.

While it is not part of any analysis or hypothesis testing, the figures below provide a descriptive visual aid of the comparison of the mean pupil size changes from baseline (and the standard deviation) for the two participant groups, monolingual vs. bilingual infants (see Figure 1), and comparisons between the three blocks, the nine pre-switch trials, the nine post-switch trials, and the 18 association trials (see Figure 2).

Figure 1. Comparison of the two groups' pupil size changes from baseline for each block.

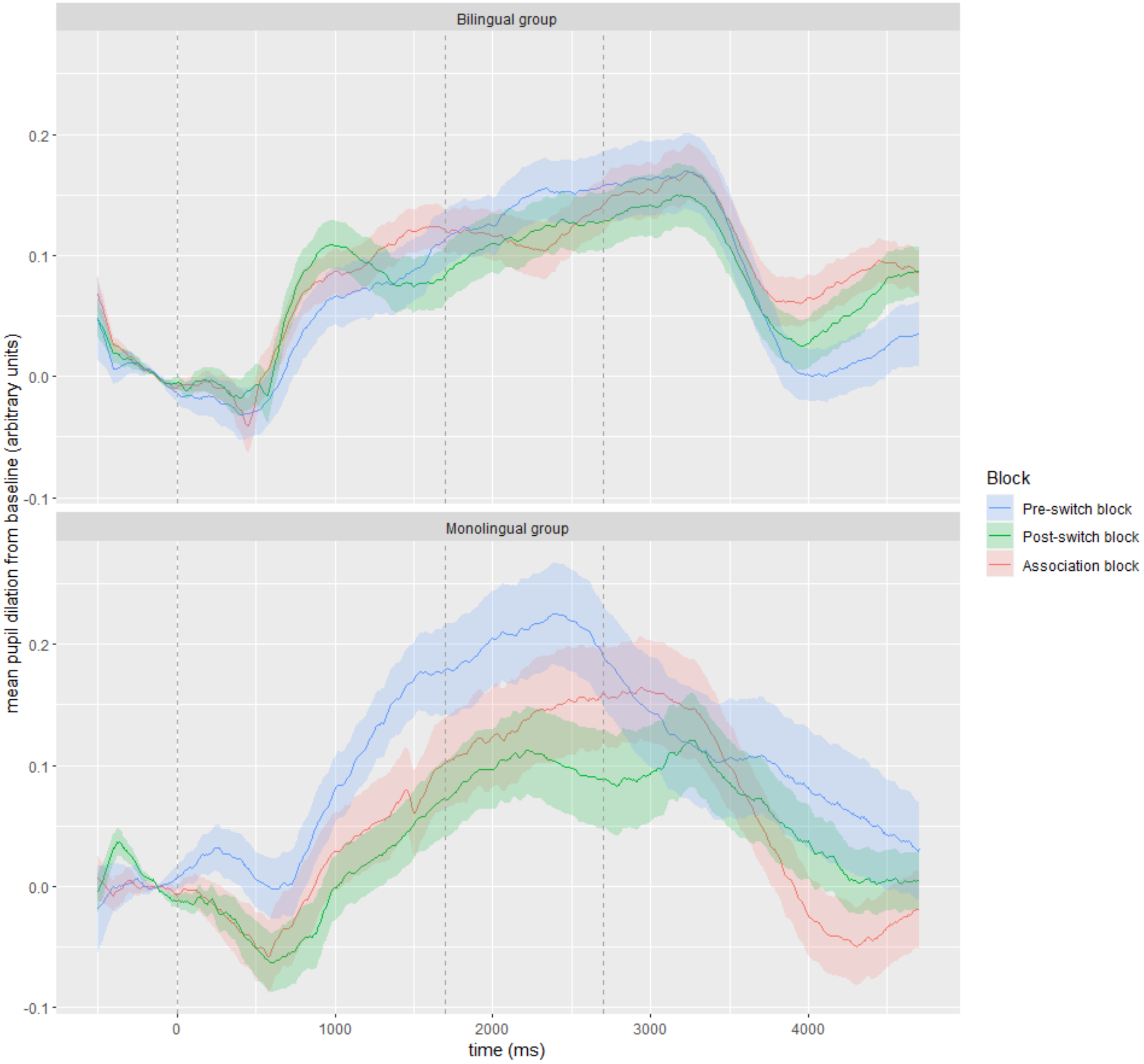


Note. Lines denote mean pupil dilation from baseline in arbitrary units[11]. The bands around the line denote the standard deviation (SD). The three dashed lines delineate the start of the stimulus phase, the anticipation phase, and the reward phase, respectively. The phase from –

[11] Some eye-trackers turn pupil size into numbers without a specific unit, thereby simply reflecting baseline size and deviations thereof, as exact distance between the eyes and the eye-tracker is not perfectly known. It is common in the literature to use this 'arbitrary unit' as a unit of measurement. This should not matter, as long as this unit remains internally consistent per participant.

500 ms to 0 ms is the attention grabber phase, which was used for calculating the baseline pupil size for each trial.

Figure 2. Comparison of pupil size changes from baseline in each block by group.



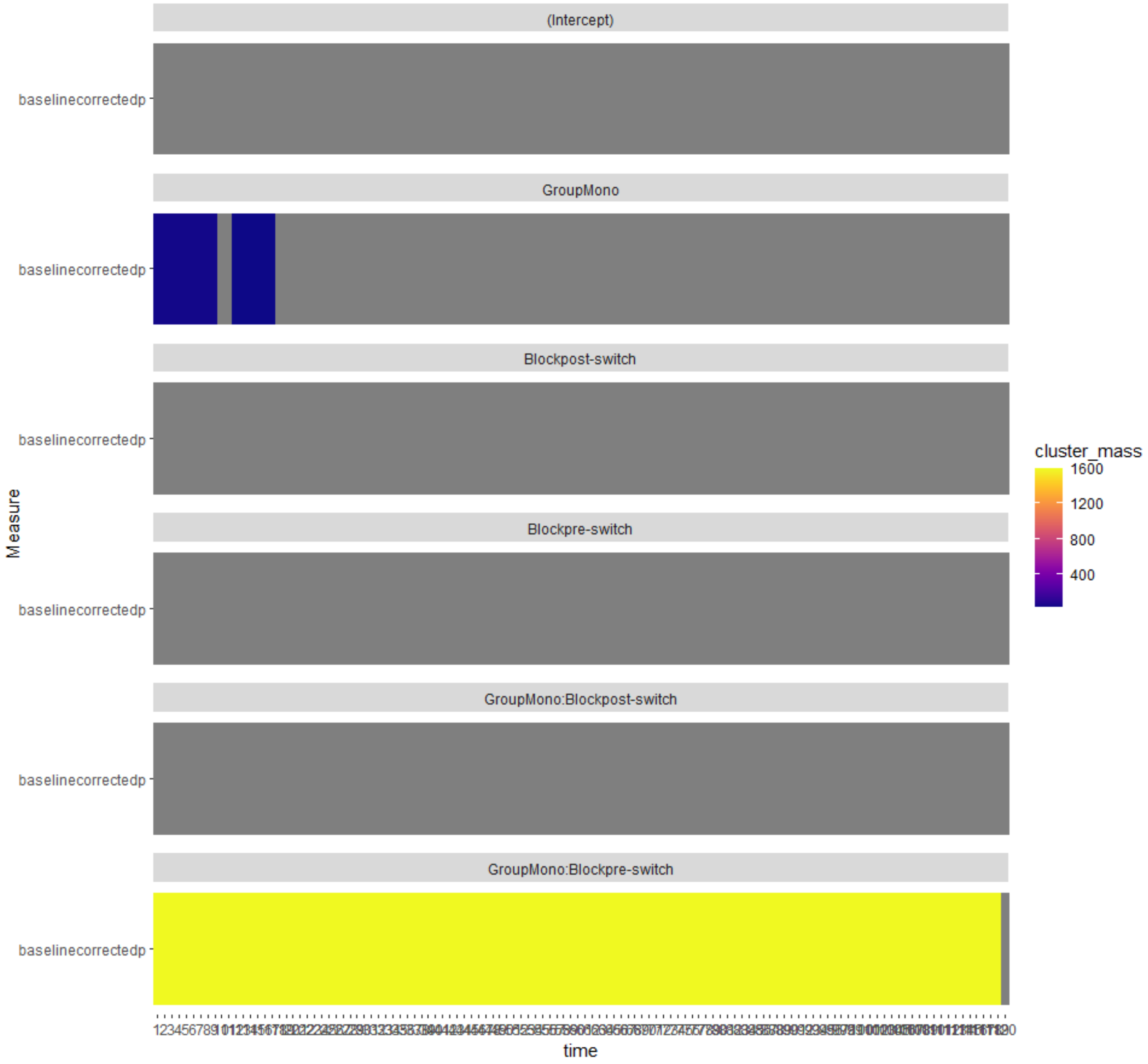
Note. Lines denote mean pupil dilation from baseline in arbitrary units, the bands around it denote the standard deviation (SD). The three dashed lines delineate the start of the stimulus phase, the anticipation phase, and the reward phase, respectively.

Anticipation Phase

The model for the entirety of the anticipation phase (1700 ms – 2700 ms) did not lead to significant results for any of the included effects. There were no main effects for group (β=.13, 95% CI [-.19, .16], *p* = .86), pre-switch vs. association block (β=.01, 95% CI [-.08, .10], *p* = .80), post-switch vs. association block (β=-.003, 95% CI [-.09, .08], *p* = .94), or interactions between group and block (monolingual x pre-switch: β=.11, 95% CI [-.04, .26], *p* = .14), (monolingual x post-switch: β=.01, 95% CI [-.12, .15], *p* = .84)). A table with all the results for the computed model can be found in Appendix A.

A permutation test was conducted to determine whether a particular time window of interest could be found within the anticipation phase (see 'GroupMono' in Figure 3). A time window of significance (cluster mass with *p* < .05) was found between 1708 and 1841 ms for the group variable, which corresponds to the start of the anticipation phase. Additionally, the entirety of the phase was shown as interesting for the monolingual group-pre-switch block interaction.

Figure 3. Results of the cluster permutation test of the anticipation phase.



Note. The bars represent the time axis for each of the main effects and interactions thereof, as labelled above the bars. Results indicate statistically significant cluster mass[12] of pupil size values: colours represent statistically significant values, with yellow blocks having higher significance than blue ones. The interaction effect of group x pre-switch block is thus highly

[12] Cluster mass can be interpreted as the probability that, under the null hypothesis, a cluster (many datapoints close together, such as pupil size difference from baseline) of statistically significant values from the mean would appear together. Under the null-hypothesis, values could deviate from the mean but this would happen at random. A cluster of deviating values in a particular time window is therefore of interest.

significant during the entirety of the phase, but specifically the group variable at the start of the phase (1708-1841 ms) is of interest as well. The time frame on the x-axis below is in arbitrary units, which were converted back to ms in the analysis.

The same mixed effects model was computed for the time window that came out as significant from the cluster permutation analysis at the group level (1708 – 1841 ms) to examine whether the independent variables led to significant differences when these were calculated for only this window of interest. However, this unfortunately did not lead to any significant results. The interaction effect closest to significance was the group and pre-switch block interaction (β = .12, 95% CI [-.01, .25], p = .064). This appeared to be the most promising outcome, yet these results were not different enough from before to justify a more precise examination.

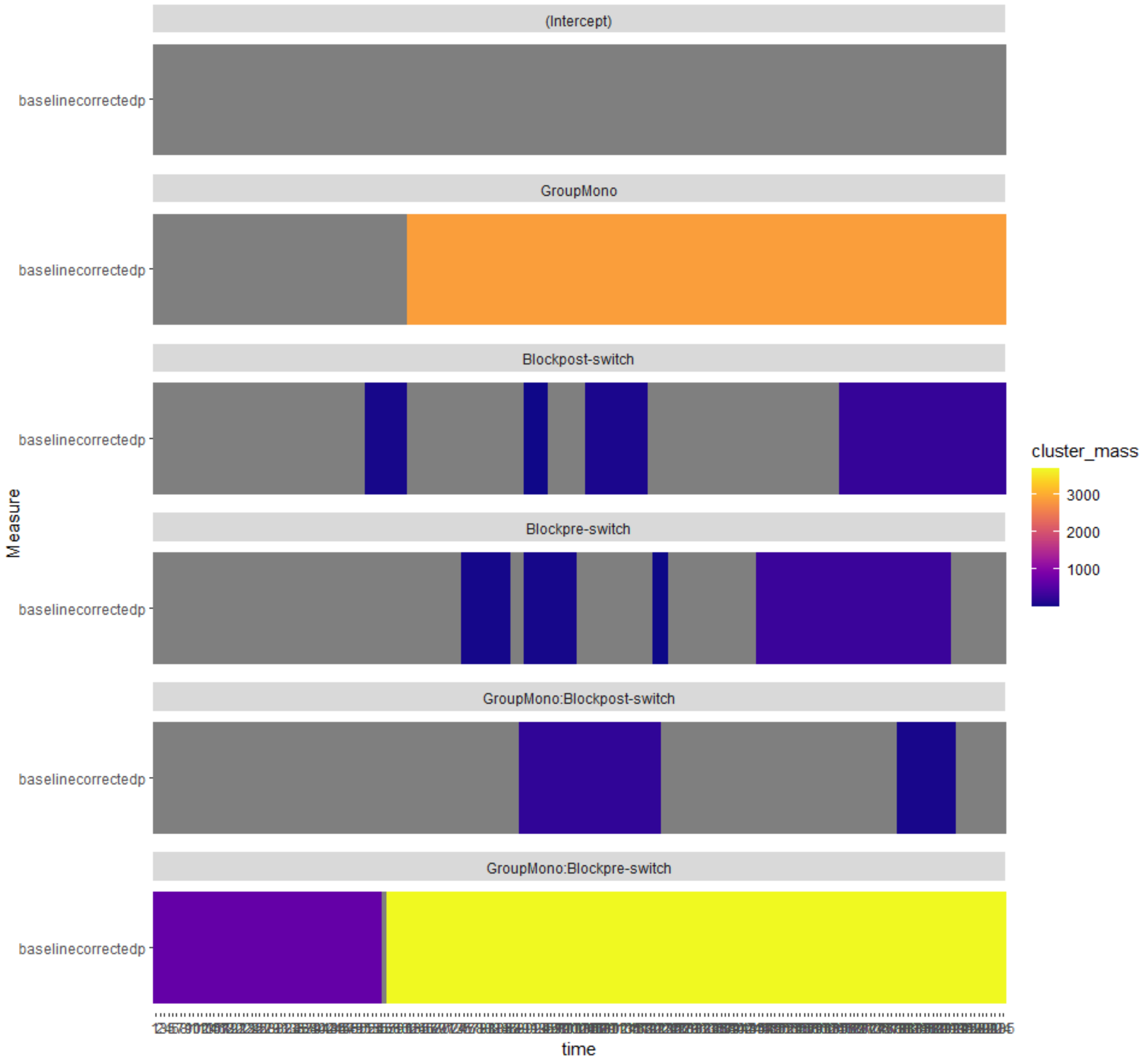
Exploratory Analysis of the Stimulus Phase

While the modelling for the anticipation phase led to null results, the other phases were also analysed using the same model as used in the anticipation phase, namely *mean_pupil ~ Group*Block + (1+Block|subject) + (1|SyllablePattern)*. The stimulus phase is the first phase of the trial, lasting from 0 to 1700 ms. In this trial, most of the effects did not reach statistical significance. However, one interaction effect led to statistically significant results: the interaction effect between group and pre-switch block was significantly different (β =.09, 95% CI [.01, 0.17], p = .034). For the full statistical results for the computed model, see Table 2 in Appendix B. Using the *phia* package for post-hoc testing, a Chi-squared test for block across group was performed while using Holm's method for p-value adjustment. The difference between the pre-switch block vs. the post-switch block across groups was statistically significant (0.11, χ^2 (1) = 6.86, p = .026), indicating that monolingual infants have a significantly greater pupil dilation from baseline in the stimulus phase of the pre-switch block (the first nine trials of the experiment) compared to the post-switch block (see Figure 2).

Another cluster permutation test was performed for this phase to find a time window of interest and see if, with a more precise time window of interest, other effects add to the model (see Figure 4). A large cluster of significance (p < .05) appeared from 508 ms after the start of the trisyllabic pattern[13], continuing for the rest of the stimulus phase (see the orange and yellow bands in Figure 4 for *GroupMono* and *GroupMono:BlockPre-switch*). Particular effects of interest are group as a main effect and an interaction for group and the pre-switch phase, as also follows from the analysis of the entire stimulus phase (see Figure 4, orange and yellow bands in the GroupMono and GroupMono:Blockpre-switch levels).

[13] This time window starting ~500 ms after the start of stimulus presentation could be related to a post-stimulus-onset delay in a PDR of ~500 ms on a cognitive task (Winn et al., 2018). See the discussion section and the corresponding footnote for further information on this.

Figure 4. Results of the cluster permutation test of the stimulus phase.



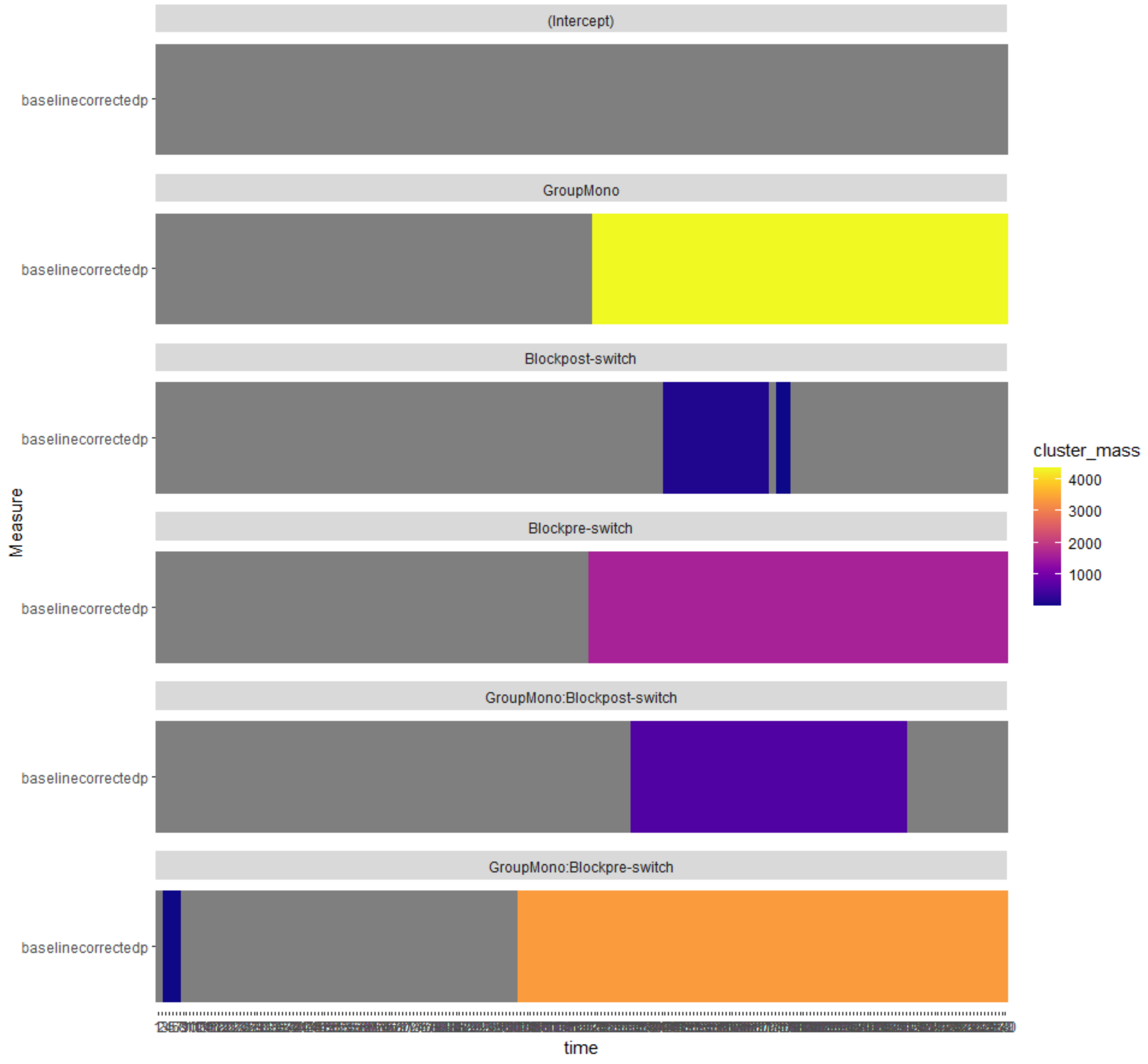
Note. The bars represent the time axis for each of the main effects and interactions thereof, as labelled above the bars. Results indicate statistically significant cluster mass of pupil size values: colours represent statistically significant values, with yellow and orange blocks having higher significance than purple and blue ones. The time frame on the x-axis below is in arbitrary units, which were converted back to ms in the analysis.

Running the same model for this selected time window alone (508 ms – 1700 ms) leads to similar results compared to the mixed effects model computed for the entire phase. Most results are not significant, apart from the interaction of group by pre-switch block (β = .10, 95% CI [0.01, 0.20], *p* = .037; see also Appendix B). A post-hoc analysis showed that the interaction over group was significant when comparing the pre-switch block to the post-switch block (0.13, χ^2 (1) = 7.17, *p* = .022), indicating once again that monolinguals have greater pupil dilations from baseline in the stimulus phase of the pre-switch block compared to bilinguals, now specifically in the 508 – 1700 ms time window. This time window in the stimulus phase appears to drive the significant interaction effect found.

Exploratory Analysis of the Reward Phase

Finally, the same exploratory analyses were run on the reward phase of the trials. The model used for all analyses, *mean_pupil ~ Group*Block + (1+Block|subject) + (1|SyllablePattern),* was also applied to the entire reward phase. None of the effects, neither main effects nor interactions thereof, were of statistical significance (see Table 3 in Appendix C). A cluster permutation test with the specified model was also performed for this phase, leading to a time window of interest from 3558 ms to the end of the trial (4700 ms) when checking for clusters at *p* < .05 (see Figure 5, coloured bands indicate time window of interest). This time window does not correspond to an event in the trial.

Figure 5. Results of the cluster permutation test of the reward phase.



Note. The bars represent the time axis for each of the main effects and interactions thereof, as labelled above the bars. Results indicate statistically significant cluster mass of pupil size values: colours represent statistically significant values, with yellow and orange blocks having higher significance than purple and blue ones. The time frame on the x-axis below is in arbitrary units, which were converted back to ms in the analysis.

The same model was once again computed for the 3558 – 4700 ms time window within the reward phase. Unlike the model for the entire reward phase, the model for the selected time window did appear to have a significant result. The interaction effect between group and block was significant (β = .13, 95% CI [02, .24], *p* = .024) (see also Appendix C). However, the results of the post-hoc analysis across group did not show a statistically significant effect between any specific blocks: the interaction effect closest to significance was the one between the pre-switch and the association blocks (0.13, χ^2 (1) = 5.11, *p* = .071). Therefore, none of the main effects or their interactions can be considered to have a statistically significant influence on pupil dilation from baseline in the reward phase.

Discussion

This pilot study is an examination of whether there is physiological evidence for a cognitive advantage in 7-month-old bilingual infants in the context of a task where attention had to be redirected and an old rule had to be inhibited. There was no evidence that there is a difference between pupil dilation responses between the bilingual infants and the monolingual control group in the post-switch block where the previously learned reward side had to be inhibited. Therefore, there is no suggestion that there is a cognitive processing advantage in bilingual infants in the context of a replication of the widely cited study by Kovács & Mehler (2009). This is in line with the results of the recent multi-centre replication study that used linear mixed models and Bayesian statistics to analyse the infants' anticipatory gaze on the same task (Spit et al., 2023). The current study used a subset of the participants from this replication study for a pupillometric analysis. Just like the behavioural replication, pupil size changes across the trial—used as an indirect physiological measurement of cognitive processing load—do not provide evidence for an advantage in pre-verbal bilinguals.

During an exploratory analysis, however, one statistically significant result was found in the linear mixed model applied to the *stimulus* phase of the trial. More specifically, a cluster permutation analysis points to a window of interest 508 ms after stimulus onset until the end of the phase, corresponding perhaps to a post-stimulus-onset delay in a PDR of ~500 ms on a cognitive task[14] (Winn et al., 2018) or ~400 ms post-onset PDR when rating music (Gingras et al., 2015). The model shows that there is an interaction between group and block: monolingual infants had a significantly larger PDR in the pre-switch block than in the post-switch block. This tentatively suggests that monolinguals have a higher cognitive load when listening to the auditorily-presented syllable pattern in the pre-switch phase. It was hypothesised that differences in the pupil dilation response were to be found in the anticipation phase of the trial, right before infants had to predict on which side the reward would show up. Additionally, it was expected that larger increases in pupil size from baseline would be found in the post-switch block compared to the pre-switch block due to requiring more effort to inhibit the previously learned rule. Monolinguals show the opposite pattern during the presentation of the syllable pattern. It is unclear what the cause of this unexpected result is. A straightforward explanation may be that the low participant number in this study leads to unreliable statistical outcomes.

Tentatively, one could also argue for a more theoretical explanation. The increased PDR in the pre-switch block compared to the post-switch block could be explained by the novelty of the stimuli: when the infants are still unfamiliar with the experiment, simply presenting them with the syllable patterns will cause an ‘orienting response’ to this sudden and unexpected event (Mathôt, 2018), which can also be seen as a pupillary novelty effect (Hepach & Westermann, 2016). This pronounced response to the presentation of the auditory

[14] Though other studies report earlier pupil responses: participants listing to snippets of music had PDRs 100-300 ms post-stimulus onset (Jagiello et al., 2019), a 300 ms post-stimulus-onset latency was found in guinea pigs being exposed to auditory stimuli (Montes-Lourido et al., 2021), and pupillary light reflexes (PLRs) are observed to have a shorter latency after light is shone in the eyes: constriction occurs at least 200 ms after the stimulus (Ellis et al., 1981).

stimulus will reduce in later blocks once the participants become more familiar with the trials or become fatigued (McLaughlin et al., 2023). However, that does not explain why an interaction effect across groups is found, with only monolinguals showing this larger pupil size change in the pre-switch block. Bilinguals appear unaffected, though there is no theoretical explanation for this. The syllables used as auditory stimuli are found in both Dutch and English, the languages the majority of infants in this study were exposed to. Thus, it is expected that both monolinguals and bilinguals would be equally as familiar with them. There are no studies (known to the author) that show that monolinguals show a novelty effect of a greater magnitude when exposed to new visual or (linguistic) auditory inputs. A study to examine whether monolingual and bilingual infants differ in this respect would help interpret the results of the current study. Below follow some suggestions as to how this could be studied.

Whether monolingual infants have a stronger pupillary response when exposed to sudden, novel, auditory or linguistic stimuli should not be difficult to study in an experimental setting: a future experiment could expose monolingual and bilingual infants to strings of unfamiliar stimuli to the children (linguistic or non-linguistic) for a prolonged period of time and see if it elicits a greater PDR for monolinguals at the beginning, while also examining how long it takes for PDRs to attenuate in both groups. Additionally, oddball paradigms have already successfully been used with pupillometry (Renner & Włodarczak, 2017), such as in a picture-word match-mismatch task, and also to examine the relationship between L2 English proficiency and the ability to discriminate between /l/ and /r/ in Japanese people (Kinzuka et al., 2020). Therefore, an auditory oddball paradigm could also be used to test monolingual and bilingual participants' PDR magnitudes across trials with habituated and novel stimuli. Whether monolingual infants have a stronger pupillary response when exposed to sudden or novel stimuli should not be difficult to study in an experimental setting: a future experiment could expose monolingual and bilingual infants to strings of unfamiliar stimuli (linguistic or non-linguistic) to children for a prolonged period of time and see if it elicits a greater PDR for monolinguals at the beginning, while also examining how long it takes for PDRs to attenuate in both groups. Additionally, oddball paradigms have already successfully been used with pupillometry (Renner & Włodarczak, 2017), such as in a picture-word match-mismatch task and also to examine the relationship between L2 English proficiency and the ability to discriminate between /l/ and /r/ in Japanese people (Kinzuka et al., 2020). Therefore, an auditory oddball paradigm could also be used to test monolinguals and bilinguals' PDR magnitudes across trials with habituated and novel stimuli.

There have also been some studies reporting a difference in attentional strategies between monolinguals and bilinguals: bilinguals sometimes appear to have a more 'exploratory' (i.e. alternating or divided) attentional strategy, where they remain more open to novel inputs and can update the information in their working memory more easily than monolinguals (see Chung-Fat-Yim et al. (2022) for a review on bilingualism and the different attention strategies). This is also mentioned as a potential mechanism for differences between monolinguals and bilinguals in some of the replication studies discussed in the 'Background Literature' section (Dal Ben et al., 2022; D'Souza et al., 2020). This state of exploratory attention[15] has been associated with a larger pupil size at baseline and smaller task-evoked PDRs, also referred to as 'tonic mode' (Joshi & Gold, 2020; Mathôt, 2018). It contrasts with the 'phasic mode' of exploitative (i.e. task-focused) attention reflected through small baseline

[15] The interplay between exploration and exploitation as two different attentional modes of behaviour, though likely not entirely distinct (Mathôt, 2018) is also known as adaptive-gain theory. It appears to be consistently associated with locus coeruleus (LC) activity (see 'Background Literature') and thus, pupil modes.

pupil sizes and larger task-evoked PDRs[16]. Perhaps something about bilinguals' tendency for exploratory attention compared to monolinguals may attenuate a measurable novelty effect in this group in the first block. However, there is currently not enough evidence to substantiate this explanation. Therefore, the reason for this outcome remains unclear and should be further investigated in future studies.

Limitations

It must be noted that the results of this paper are subject to some methodological and statistical limitations: the sample size of both participant groups was small due to COVID-related recruitment difficulties and the rejection of participants due to distractedness. As such, only nine bilingual infants and six monolingual infants were included in the data analysis, which also leads to the bilingual group being 1.5 larger in size. Additionally, as advanced statistical methods were beyond the scope of this paper and more complex modelling efforts (e.g. GAMMs and LMMs for the full trial with more independent variables) led to non-convergence issues and failed model assumption tests (e.g. the residuals of models were non-normally distributed), a simpler mixed model was used than initially intended. Additionally, the frequentist approach of significance testing was used to test for statistical significance instead of using the Bayesian approach to significance testing, employed in Spit et al. (2023). Unlike the Bayesian approach, the frequentist approach to significance testing does not give the probability ratio of the alternative hypothesis over the null hypothesis given the data (Johnson, 2013); it acquires point estimates for values (such as pupil size) and bases the

[16] Decreased attention in general is marked by both a small baseline pupil size and attenuated PDRs (Mathôt, 2018).

significance of this point estimate on the hypothetical percentage of samples (i.e. data sets) containing the same or a more extreme distribution of data in a scenario where the same experiment was repeated *ad infinitum* and the null hypothesis would be true. A *p*-value, however, does not claim anything about the alternative hypothesis.

It is certainly possible that the significant PDR in the pre-switch block for monolinguals alone could be a statistical artefact related either to the limitations mentioned above or simply introduced due to chance. Moreover, as the only significant result was found in an exploratory analysis of the stimulus phase, no definitive conclusions can be drawn.

There are also several methodological limitations in the study design and procedure: For example, the bilinguals used in the current study were unbalanced bilinguals[17], who often had more exposure to one language than to another (mean relative L1 exposure = 66.5%, SD = 8.3%, range = 55 – 76.3%). Language exposure for all participants was also reported by the caregivers through an estimation, so there is an unknown rate of error to this. Furthermore, quality and type of linguistic input were not described (e.g. was the infant directly spoken to in a language or was a large part of the input caregivers talking in proximity of the child?). It is therefore possible that some bilinguals have a lot less non-dominant language input than other bilinguals. It could be argued that this is relevant to their expected performance during the experiment, as they could have less experience with "switching" languages.

Another issue is that pupil size fluctuations due to the pupillary light reflex cannot be ruled out: while ambient lighting was kept as consistent as possible during the experiment and

[17] The bilingual participants in Spit et al. (2023) were also unbalanced, though with a maximum relative exposure of 75% input in the dominant language, whereas the current study uses a maximum of 80% linguistic input in the L1. The relative linguistic input for the languages in the bilingual group of the original Kovács and Mehler (2009) study are unknown.

across participants, the stimuli used on the screen (i.e. the reward showing up in the white boxes) were not designed with consistency in their luminosity in mind, as these were part of an exact replication of experiment 2 of Kovács and Mehler (2009). However, the significant results that were found did not correspond to the time windows of the trial with the reward pictures showing up, and the white boxes remained on the screen for the entire duration of the experiment.

Perhaps another limitation is the task complexity of the current study design and thus may ask too much of the young participants. More distinctive pre-reward stimuli (e.g. entirely red or green visual cues instead of slight syllable pattern differences), not linking specific stimuli to a reward side (as done by Dal Ben et al. (2022)), or longer anticipation phases for prediction and longer breaks between trials may contribute to clearer results. Other experimental paradigms could be employed to see if different contexts lead to a reliable difference in pupillometric responses between groups, or perhaps even different task difficulty levels, as this affects pupil dilation responses (Dunst et al., 2014). Furthermore, the experiment was twice as long as the original by Kovács and Mehler (2009) due to the addition of the 18 association trials. The infants participating in the experiment may not be able to stay alert and focus on the experiment throughout the duration of the trials. Fatigue leads to decreases in pupil size fluctuations (Sirois & Brisson, 2014), so task-related PDRs could be attenuated when infants disengage from the experiment (McLaughlin et al., 2023).

The previous point also relates to a gap in the current statistical analysis: baseline pupil size of the participants per trial could be informative to infer the infant's attentional strategy during the experiment, where ‘tonic mode’ (see above), as determined by a larger baseline pupil size, is associated with an exploratory attentional strategy, in contrast with the ‘phasic mode’, associated with lower baseline values yet still notable task-related PDRs, indicating an exploitative attentional strategy (Joshi & Gold, 2020; Mathôt, 2018). Baseline pupil size and peak PDRs from baseline per trial could be used as a first exploratory analysis in order to examine an infant's attentional strategy per trial and to test whether an infant is still paying attention to the experiment or is fatigued, as fatigue is indicated by a small baseline pupil size and attenuated PDRs (see above). This exploratory analysis could be done in the future on the current dataset or, preferably, on a replication with more participants.

Another limitation relates to the eye-tracker itself: the Tobii-T120 has been found to have a measurement error in pupil diameter when the participant is looking away from the middle of the screen: looking in another direction than straight on causes the pupil to become slightly flattened from the eye-tracker’s perspective (Brisson et al., 2013). The set-up itself may have led to measurement inconsistencies as well, as infants were sitting on their caregiver’s lap in front of the eye-tracker’s screen. Thus, they had more freedom to move their head around than with eye-tracking glasses or when using a set-up with head support, meaning they did not always retain the exact 60 cm distance from the screen during the experiment. While this issue affects outcomes less with using a per-trial pupil size baseline and similar infant pupillometry experiment set-ups are previously used with a Tobii T120 eye-tracker (see e.g. Jackson & Sirois, 2009; 2022), it cannot be fully ruled out that this may have had some effect.

From a more fundamental perspective, it must be noted that pupil size fluctuations, while linked to cognitive load[18], cannot determine the exact reason (i.e. specific cognitive process) for the increased cognitive load. This is both a blessing and a curse, as this means PDRs can be used as a general substrate for multiple proposed cognitive processing mechanisms mentioned in the literature about the bilingual cognitive advantage. However, the downside is that we therefore cannot make any theoretical claims as to which specific processes actually

[18] That is, if no fluctuations in environmental light sources occur, as that will lead to a pupillary light reflex.

lead to a PDR. PDRs can therefore be used to investigate whether there are cognitive processing differences between groups or contexts, not whether these differences relate to specific processes related to attention, inhibitory control, cognitive arousal or expectations, to name a few. More informative measurements regarding which neural networks are activated can be provided with fNIRS or fMRI set-ups.

Due to the limitations in study design, sample size and statistical methodology, the current study should mostly be considered as a proof-of-concept study to inform one of the possibilities and limitations of using pupillometry in infant cognition and bilingualism research. While shortcomings apply, pupillometry may provide value in investigating the presence or absence of a cognitive advantage in the bilingual population. So far, this pilot study is consistent with recent null findings related to the bilingual advantage in infants within the same experimental paradigm (D'Souza et al., 2020; Kalashnikova et al., 2021; Spit et al., 2023). Thus, it is in accordance with the literature debating the robustness and reliability of the bilingual cognitive advantage (de Bruin et al., 2014, 2021; Duñabeitia et al., 2014; Hernández et al., 2013; Jones et al., 2021; Paap et al., 2015; Paap & Greenberg, 2013), though future studies with larger sample sizes may lead to different results.

References

Addyman, C., Rocha, S., & Mareschal, D. (2014). Mapping the origins of time: Scalar errors in infant time estimation. *Developmental Psychology*, *50*(8), 2030. https://doi.org/10.1037/A0037108

Adesope, O. O., Lavin, T., Thompson, T., & Ungerleider, C. (2010). A Systematic Review and Meta-Analysis of the Cognitive Correlates of Bilingualism. *The Review of Educational Research*, *80*(2), 207–245. https://doi.org/10.3102/0034654310368803

Anderson, J. A. E., Hawrylewicz, K., & Grundy, J. G. (2020). Does bilingualism protect against dementia? A meta-analysis. *Psychonomic Bulletin and Review*, *27*(5), 952–965. https://doi.org/10.3758/S13423-020-01736-5/

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278. https://doi.org/10.1016/J.JML.2012.11.001

Bates, D., Kliegl, R., Vasishth, S., & Baayen, R. H. (2015). Parsimonious Mixed Models. https://arxiv.org/abs/1506.04967v2

Bates, D., Mächler, M., Bolker, B. M., & Walker, S. C. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/JSS.V067.I01

Bialystok, E. (2008). Cognitive Effects of Bilingualism across the Lifespan. *Proceedings of the Annual Boston University Conference on Language Development*, *32*(1).

Bialystok, E. (2009). Bilingualism: The good, the bad, and the indifferent*. *Bilingualism: Language and Cognition*, *12*(1), 3–11. https://doi.org/10.1017/S1366728908003477

Bialystok, E., & Craik, F. I. M. (2022). How does bilingualism modify cognitive function? Attention to the mechanism. *Psychonomic Bulletin & Review 2022 29:4*, *29*(4), 1246–1269. https://doi.org/10.3758/S13423-022-02057-5

Bialystok, E., Craik, F. I. M., & Freedman, M. (2007). Bilingualism as a protection against the onset of symptoms of dementia. *Neuropsychologia*, *45*(2), 459–464. https://doi.org/10.1016/J.NEUROPSYCHOLOGIA.2006.10.009

Bialystok, E., Craik, F. I. M., Grady, C., Chau, W., Ishii, R., Gunji, A., & Pantev, C. (2005). Effect of bilingualism on cognitive control in the Simon task: Evidence from MEG. *NeuroImage*, *24*(1). https://doi.org/10.1016/j.neuroimage.2004.09.044

Bosma, E., & Pablos, L. (2020). Switching direction modulates the engagement of cognitive control in bilingual reading comprehension: An ERP study. *Journal of Neurolinguistics*, *55*, 100894. https://doi.org/10.1016/J.JNEUROLING.2020.100894

Brisson, J., Mainville, M., Mailloux, D., Beaulieu, C., Serres, J., & Sirois, S. (2013). Pupil diameter measurement errors as a function of gaze direction in corneal reflection eyetrackers. *Behavior Research Methods*, *45*(4), 1322–1331. https://doi.org/10.3758/S13428-013-0327-0/

Chung-Fat-Yim, A., Calvo, N., & Grundy, J. G. (2022). The Multifaceted Nature of Bilingualism and Attention. *Frontiers in Psychology*, *13*, 910382. https://doi.org/10.3389/FPSYG.2022.910382/

Dal Ben, R., Killam, H., Pour Iliaei, S., & Byers-Heinlein, K. (2022). Bilingualism Affects Infant Cognition: Insights From New and Open Data. *Open Mind*, *6*, 88–117. https://doi.org/10.1162/OPMI_A_00057

Dalmaijer, E. S., Mathôt, S., & Van der Stigchel, S. (2014). PyGaze: an open-source, cross-platform toolbox for minimal-effort programming of eyetracking experiments. *Behavior Research Methods*, *46*(4), 913–921. https://doi.org/10.3758/S13428-013-0422-2/

De Bruin, A., Dick, A. S., & Carreiras, M. (2021). Clear Theories are Needed to Interpret Differences: Perspectives on the Bilingual Advantage Debate. *Neurobiology of Language*. https://doi.org/10.1162/nol_a_00038

De Bruin, A., Treccani, B., & della Sala, S. (2014). Cognitive Advantage in Bilingualism: An Example of Publication Bias? *Psychological Science*, *26*(1), 99–107. https://doi.org/10.1177/0956797614557866

De Vries, L. M., Amelynck, S., Nyström, P., van Esch, L., Van Lierde, T., Warreyn, P., Roeyers, H., Noens, I., Naulaers, G., Boets, B., Steyaert, J., Moerman, F., Erdogan, M., Mađarević, M., & Segers, J. (2023). Investigating the development of the autonomic nervous system in infancy through pupillometry. *Journal of Neural Transmission*, *130*(5), 723. https://doi.org/10.1007/S00702-023-02616-7

Di Domenico, S. I., Rodrigo, A. H., Ayaz, H., Fournier, M. A., & Ruocco, A. C. (2015). Decision-making conflict and the neural efficiency hypothesis of intelligence: A functional near-infrared spectroscopy investigation. *NeuroImage*, *109*, 307–317. https://doi.org/10.1016/J.NEUROIMAGE.2015.01.039

Donnelly, S., Brooks, P. J., & Homer, B. D. (2015). Examining the Bilingual Advantage on Conflict Resolution Tasks: A Meta-Analysis. *37th Annual Conference of the Cognitive Science Society*.

D'Souza, D., Brady, D., Haensel, J. X., & D'Souza, H. (2020). Is mere exposure enough? The effects of bilingual environments on infant cognitive development. *Royal Society Open Science*, *7*(2). https://doi.org/10.1098/RSOS.180191

Duñabeitia, J. A., Hernández, J. A., Antón, E., Macizo, P., Estévez, A., Fuentes, L. J., & Carreiras, M. (2014). The inhibitory advantage in bilingual children revisited: Myth or reality? *Experimental Psychology*, *61*(3). https://doi.org/10.1027/1618-3169/a000243

Dunst, B., Benedek, M., Jauk, E., Bergner, S., Koschutnig, K., Sommer, M., Ischebeck, A., Spinath, B., Arendasy, M., Bühner, M., Freudenthaler, H., & Neubauer, A. C. (2014). Neural efficiency as a function of task demands. *Intelligence*, *42*(1), 22. https://doi.org/10.1016/J.INTELL.2013.09.005

Ellis, C. J. K., Thomas's Hospital, S., & Se, L. (1981). The pupillary light reflex in normal subjects. *British Journal of Ophthalmology*, *65*(11), 754–759. https://doi.org/10.1136/BJO.65.11.754

Forbes, S. H. (2020). PupillometryR: An R package for preparing and analysing pupillometry data. *Journal of Open Source Software*, *5*(50), 2285. https://doi.org/10.21105/JOSS.02285

Frossard, J., & Renaud, O. (2021). Permutation Tests for Regression, ANOVA, and Comparison of Signals: The permuco Package. *Journal of Statistical Software, 99*(15), 1–32. https://doi.org/10.18637/jss.v099.i15

Geller, J., Winn, M. B., Mahr, T., & Mirman, D. (2020). GazeR: A Package for Processing Gaze Position and Pupil Size Data. *Behavior Research Methods*, *52*(5), 2232–2255. https://doi.org/10.3758/S13428-020-01374-8/

Gingras, B., Marin, M. M., Puig-Waldmüller, E., & Fitch, W. T. (2015). The eye is listening: Music-induced arousal and individual differences predict pupillary responses. *Frontiers in Human Neuroscience*, *9*. https://doi.org/10.3389/FNHUM.2015.00619/

Gredebäck, G., & Melinder, A. (2011). Teleological Reasoning in 4-Month-Old Infants: Pupil Dilations and Contextual Constraints. *PLOS ONE, 6*(10), e26487. https://doi.org/10.1371/JOURNAL.PONE.0026487

Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(2), 67–81. https://doi.org/10.1017/S1366728998000133

Green, D. W., & Abutalebi, J. (2013). Language control in bilinguals: The adaptive control hypothesis. *Journal of Cognitive Psychology*, *25*(5), 515–530. https://doi.org/10.1080/20445911.2013.796377

Heise, M. J., Mon, S. K., & Bowman, L. C. (2022). Utility of linear mixed effects models for event-related potential research with infants and children. *Developmental Cognitive Neuroscience*, *54*, 101070. https://doi.org/10.1016/J.DCN.2022.101070

Hepach, R., & Westermann, G. (2016). Pupillometry in Infancy Research. *Journal of Cognition and Development*, *17*(3). https://doi.org/10.1080/15248372.2015.1135801

Hernández, M., Martin, C. D., Barceló, F., & Costa, A. (2013). Where is the bilingual advantage in task-switching? *Journal of Memory and Language*, *69*(3). https://doi.org/10.1016/j.jml.2013.06.004

Hershaw, J. N., & Ettenhofer, M. L. (2018). Insights into cognitive pupillometry: Evaluation of the utility of pupillary metrics for assessing cognitive load in normative and clinical samples. *International Journal of Psychophysiology*, *134*, 62–78. https://doi.org/10.1016/J.IJPSYCHO.2018.10.008

Hess, E. H., & Polt, J. M. (1964). Pupil Size in Relation to Mental Activity during Simple Problem-Solving. *Science*, *143*(3611), 1190–1192. https://doi.org/10.1126/SCIENCE.143.3611.1190

Hilchey, M. D., & Klein, R. M. (2011). Are there bilingual advantages on nonlinguistic interference tasks? Implications for the plasticity of executive control processes. *Psychonomic Bulletin & Review 2011 18:4*, *18*(4), 625–658. https://doi.org/10.3758/S13423-011-0116-7

Jackson, I. R., & Sirois, S. (2022). But that’s possible! Infants, pupils, and impossible events. *Infant Behavior and Development*, *67*. https://doi.org/10.1016/J.INFBEH.2022.101710

Jackson, I., & Sirois, S. (2009). Infant cognition: going full factorial with pupil dilation. *Developmental Science*, *12*(4), 670–679. https://doi.org/10.1111/J.1467-7687.2008.00805.X

Jagiello, R., Pomper, U., Yoneya, M., Zhao, S., & Chait, M. (2019). Rapid Brain Responses to Familiar vs. Unfamiliar Music – an EEG and Pupillometry study. *Scientific Reports, 9:1, 9*(1), 1–13. https://doi.org/10.1038/s41598-019-51759-9

Johnson, V. E. (2013). Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences of the United States of America*, *110*(48), 19313–19317. https://doi.org/10.1073/pnas.1313476110

Jones, S. K., Davies-Thompson, J., & Tree, J. (2021). Can Machines Find the Bilingual Advantage? Machine Learning Algorithms Find No Evidence to Differentiate Between Lifelong Bilingual and Monolingual Cognitive Profiles. *Frontiers in Human Neuroscience*, *15*. https://doi.org/10.3389/fnhum.2021.621772

Joshi, S., & Gold, J. I. (2020). Pupil Size as a Window on Neural Substrates of Cognition. *Trends in Cognitive Sciences*, *24*(6), 466–480. https://doi.org/10.1016/J.TICS.2020.03.005

Kahneman, D., & Beatty, J. (1966). Pupil Diameter and Load on Memory. *Science*, *154*(3756), 1583–1585. https://doi.org/10.1126/SCIENCE.154.3756.1583

Kalashnikova, M., Pejovic, J., & Carreiras, M. (2021). The effects of bilingualism on attentional processes in the first year of life. *Developmental Science*, *24*(2), e13011. https://doi.org/10.1111/DESC.13011

Kinzuka, Y., Minami, T., & Nakauchi, S. (2020). Pupil dilation reflects English /l//r/ discrimination ability for Japanese learners of English: a pilot study. *Scientific Reports, 10:1*, *10*(1), 1–9. https://doi.org/10.1038/s41598-020-65020-1

Kovács, Á. M. (2009). Early bilingualism enhances mechanisms of false-belief reasoning. *Developmental Science*, *12*(1). https://doi.org/10.1111/j.1467-7687.2008.00742.x

Kovács, Á. M., & Mehler, J. (2009). Cognitive gains in 7-month-old bilingual infants. *Proceedings of the National Academy of Sciences*, *106*(16), 6556–6560. https://doi.org/10.1073/PNAS.0811323106

Kret, M. E., & Sjak-Shie, E. E. (2019). Preprocessing pupil size data: Guidelines and code. *Behavior Research Methods*, *51*(3), 1336–1342. https://doi.org/10.3758/S13428-018-1075-Y

Laeng, B., & Alnaes, D. (2019). *Pupillometry*. 449–502. https://doi.org/10.1007/978-3-030-20085-5_11

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A Window to the Preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

Mathôt, S. (2018). Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition*, *1*(1), 1–23. https://doi.org/10.5334/JOC.18

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, *50*(1), 94–106. https://doi.org/10.3758/S13428-017-1007-2

Mathôt, S., & Vilotijević, A. (2022). Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behavior Research Methods*, *1*, 1–23. https://doi.org/10.3758/S13428-022-01957-7/

Matuschek, H., Kliegl, R., Vasishth, S., Baayen, H., & Bates, D. (2017). Balancing Type I error and power in linear mixed models. *Journal of Memory and Language*, *94*, 305–315. https://doi.org/10.1016/J.JML.2017.01.001

McLaughlin, D. J., Zink, M. E., Gaunt, L., Reilly, J., Sommers, M. S., Van Engen, K. J., & Peelle, J. E. (2023). Give me a break! Unavoidable fatigue effects in cognitive pupillometry. *Psychophysiology*, *60*(7). https://doi.org/10.1111/PSYP.14256

Montes-Lourido, P., Kar, M., Kumbam, I., & Sadagopan, S. (2021). Pupillometry as a reliable metric of auditory detection and discrimination across diverse stimulus paradigms in animal models. *Scientific Reports, 11:1*, *11*(1), 1–15. https://doi.org/10.1038/s41598-021-82340-y

Paap, K. R., & Greenberg, Z. I. (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, *66*(2). https://doi.org/10.1016/j.cogpsych.2012.12.002

Paap, K. R., Johnson, H. A., & Sawi, O. (2015). Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. *Cortex* (Vol. 69). https://doi.org/10.1016/j.cortex.2015.04.014

Renner, L. F., & Włodarczak, M. (2017). When a dog is a cat and how it changes your pupil size: Pupil dilation in response to information mismatch. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2017-August*, 674–678. https://doi.org/10.21437/INTERSPEECH.2017-353

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. https://doi.org/10.1080/13506285.2019.1600089

R Team. (2014). R: A language and environment for statistical computing. *MSOR Connections*.

Sabourin, L., & Vinerte, S. (2015). The bilingual advantage in the Stroop task: Simultaneous vs. early bilinguals. *Bilingualism*, *18*(2). https://doi.org/10.1017/S1366728914000704

Sirois, S., & Brisson, J. (2014). Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science, 5*(6), 679–692. https://doi.org/10.1002/WCS.1323

Spit, S., Geambașu, A., Van Renswoude, D., Blom, E., Fikkert, P., Hunnius, S., Junge, C., Verhagen, J., Visser, I., Wijnen, F., & Levelt, C. C. (2023). Robustness of the cognitive gains in 7-month-old bilingual infants: A close multi-center replication of Kovács and Mehler (2009). *Developmental Science*. https://doi.org/10.1111/desc.13377

Van den Berg, F., Brouwer, J., Tienkamp, T. B., Verhagen, J., & Keijzer, M. (2022). Language Entropy Relates to Behavioral and Pupil Indices of Executive Control in Young Adult Bilinguals. *Frontiers in Psychology*, *13*, 864763. https://doi.org/10.3389/FPSYG.2022.864763/

Van Rij, J., Hendriks, P., van Rijn, H., Baayen, R. H., & Wood, S. N. (2019). Analyzing the Time Course of Pupillometric Data. *Trends in Hearing*, *23*. https://doi.org/10.1177/2331216519832483

Verney, S. P., Granholm, E., & Marshall, S. P. (2004). Pupillary responses on the visual backward masking task reflect general cognitive ability. *International Journal of Psychophysiology*, *52*(1), 23–36. https://doi.org/10.1016/J.IJPSYCHO.2003.12.003

Voeten, C. C. (2021). Analyzing time series data using clusterperm.lmer.

Winn, M. B., Wendt, D., Koelewijn, T., & Kuchinsky, S. E. (2018). Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started. *Trends in Hearing*, *22*. https://doi.org/10.1177/2331216518800869

Zhang, F., & Emberson, L. L. (2020). Using pupillometry to investigate predictive processes in infancy. *Infancy*, *25*(6). https://doi.org/10.1111/infa.12358

Zhang, F., Jaffe-Dax, S., Wilson, R. C., & Emberson, L. L. (2019). Prediction in infants and adults: A pupillometry study. *Developmental Science*, *22*(4). https://doi.org/10.1111/DESC.12780

Appendix A

Table 1. *Results of the linear mixed effects regression for the entire anticipation phase (1700 – 2700 ms), and the permutation time window (1708- 1841 ms)*

		Entire anticipation phase (1700 – 2700 ms)					Permutation time window (1708-1841 ms)				
Effects		*β*	SE	CI (95%)	*t*	*p*	*β*	SE	CI (95%)	*t*	*p*
Fixed	(Intercept)	.13	.06	[.02, .24]	2.29	.02	.13	.05	[.03, .23]	2.47	.01
	Monolingual group	-.02	.09	[-.19, .16]	-.17	.86	-.03	.08	[-.19, .13]	-.37	.71
	Pre-switch block	.01	.05	[-.08, .10]	.24	.80	-.01	.04	[-.09, .07]	-.24	.81
	Post-switch block	-.003	.04	[-.09, .08]	-.07	.94	-.03	.04	[-.11, .06]	-.59	.55
	Monolingual:pre-switch	.11	.08	[-.04, .26]	1.45	.14	.12	.07	[-.01, .25]	1.86	.06
	Monolingual:post-switch	.01	.07	[-.12, .15]	.20	.84	.03	.07	[-.10, .17]	.49	.63
Random		Var.	SD				Var.	SD			
	Syllable pattern (item)	.0007	.008				.0001	.011			
	Subject	.025	.157				.021	.144			
	Subject:pre-switch	.012	.109				.011	.104			
	Subject:post-switch	.010	.098				.008	.090			
	Residuals	.031	.176				.027	.164			

Note. β = estimate/coefficient, SE = standard error, CI = confidence interval, Var. = variance, SD = standard deviation.

Table 2. *Results of the linear mixed effects regression for the entire stimulus phase (0 – 1700 ms), and the permutation time window (508 – 1700 ms)*

		Entire stimulus phase (0 – 1700 ms)					Permutation time window (508-1700 ms)				
Effects		*β*	SE	CI (95%)	*t*	*p*	*β*	SE	CI (95%)	*t*	*p*
Fixed	(Intercept)	.05	.04	[-.02, .12]	1.32	.19	.08	.05	[-.01, .17]	1.71	.09
	Monolingual group	-.03	.06	[-.15, .09]	-.52	.60	-.05	.07	[-.19, .10]	-.62	.53
	Pre-switch block	-.02	.02	[-.07, .03]	.74	.46	-.02	.03	[-.08, .03]	-.81	.42
	Post-switch block	.005	.03	[-.05, .06]	-.19	.85	.002	.03	[-.07, .07]	.06	.95
	Monolingual:pre-switch	.09	.04	[.01, .17]	2.12	.034*	.10	.05	[.01, .20]	2.09	.037*
	Monolingual:post-switch	-.02	.05	[-.11, .07]	-.42	.68	-.02	.06	[-.13, .09]	-.39	.70
Random		Var.	SD				Var.	SD			
	Syllable pattern (item)	.0002	.015				.0003	.017			
	Subject	.011	.105				.017	.130			
	Subject:pre-switch	.002	.049				.004	.063			
	Subject:post-switch	.004	.063				.006	.080			
	Residuals	.014	.117				.019	.136			

Note. β = estimate/coefficient, SE = standard error, CI = confidence interval, Var. = variance, SD = standard deviation.

Table 3. *Results of the linear mixed effects regression for the entire reward phase (2700 – 4700 ms), and the permutation time window (3558 – 4700 ms)*

		Entire stimulus phase (2700 – 4700 ms)					Permutation time window (3558-4700 ms)				
Effects		*β*	SE	CI (95%)	*t*	*p*	*β*	SE	CI (95%)	*t*	*p*
Fixed	(Intercept)	.12	.04	[.04, .20]	2.91	.004	.09	.03	[.02, .16]	2.56	.01
	Monolingual group	-.04	.07	[-.18, .09]	-.65	.52	-.07	.06	[-.18, .04]	-1.21	.23
	Pre-switch block	-.04	.03	[-.11, .02]	-1.27	.20	-.06	.03	[-.13, .00]	-1.95	.051
	Post-switch block	-.02	.03	[-09., .04]	-.76	.45	-.03	.03	[-.09, .03]	-.92	.36
	Monolingual:pre-switch	.08	.06	[-.03, .19]	1.42	.16	.13	.06	[.02, .24]	2.26	.024*
	Monolingual:post-switch	.01	.05	[-.09, .11]	.21	.83	.04	.05	[-.06, .14]	0.73	.46
Random		Var.	SD				Var.	SD			
	Syllable pattern (item)	.0007	.026				.0008	.028			
	Subject	.012	.109				.007	.081			
	Subject:pre-switch	.002	.047				.002	.045			
	Subject:post-switch	.001	.035				.0001	.011			
	Residuals	.032	.180				.034	.185			

Note. β = estimate/coefficient, SE = standard error, CI = confidence interval, Var. = variance, SD = standard deviation.