Infant Behavior and Development 73 (2023) 101890

Contents lists available at ScienceDirect

Infant Behavior and Development

journal homepage: www.elsevier.com/locate/inbede

The pupil collaboration: A multi-lab, multi-method analysis of goal attribution in infants

Sylvain Sirois [a,*], Julie Brisson [b], Erik Blaser [c], Giulia Calignano [d], Jamie Donenfeld [c], Robert Hepach [e], Jean-Rémy Hochmann [f], Zsuzsa Kaldy [c], Ulf Liszkowski [g], Marlena Mayer [g], Shannon Ross-Sheehy [h], Sofia Russo [d], Eloisa Valenza [d]

[a] *Département de Psychologie, Université du Québec à Trois-Rivières, Canada*
[b] *Centre de Recherche sur les fonctionnements et dysfonctionnements psychologiques (EA7475), Université de Rouen Normandie, France*
[c] *Department of Psychology, University of Massachusetts Boston, USA*
[d] *Department of Developmental and Social Psychology, University of Padova, Italy*
[e] *Department of Experimental Psychology, University of Oxford, UK*
[f] *CNRS UMR5229 - Institut des Sciences Cognitives Marc Jeannerod, Université Lyon 1, France*
[g] *Department of Developmental Psychology, University of Hamburg, Germany*
[h] *Department of Psychology, University of Tennessee, USA*

ARTICLE INFO

Keywords:
Goal attribution
Infancy
Pupil dilation

ABSTRACT

The rise of pupillometry in infant research over the last decade is associated with a variety of methods for data preprocessing and analysis. Although pupil diameter is increasingly recognized as an alternative measure of the popular cumulative looking time approach used in many studies (Jackson & Sirois, 2022), an open question is whether the many approaches used to analyse this variable converge. To this end, we proposed a crowdsourced approach to pupillometry analysis. A dataset from 30 9-month-old infants (15 girls; M_{age} = 282.9 days, SD = 8.10) was provided to 7 distinct teams for analysis. The data were obtained from infants watching video sequences showing a hand, initially resting between two toys, grabbing one of them (after Woodward, 1998). After habituation, infants were shown (in random order) a sequence of four test events that varied target position and target toy. Results show that looking times reflect primarily the familiar path of the hand, regardless of target toy. Gaze data similarly show this familiarity effect of path. The pupil dilation analyses show that features of pupil baseline measures (duration and temporal location) as well as data retention variation (trial and/or participant) due to different inclusion criteria from the various analysis methods are linked to divergences in findings. Two of the seven teams found no significant findings, whereas the remaining five teams differ in the pattern of findings for main and interaction effects. The discussion proposes guidelines for best practice in the analysis of pupillometry data.

* Correspondence to: Département de Psychologie Université du Québec à Trois-Rivières 3351, boulevard des Forges Trois-Rivières (Québec) G8Z 4M3 Canada.
E-mail address: sylvain.sirois@uqtr.ca (S. Sirois).

https://doi.org/10.1016/j.infbeh.2023.101890
Received 9 June 2022; Received in revised form 27 September 2023; Accepted 27 September 2023
Available online 7 November 2023

1. Introduction

1.1. Replication strategies

Replicability can be construed as a fundamental, grounding property of science (Frank et al., 2023; Romero, 2019). Behavioral research can be trusted when a new finding can be reproduced using the same methods with similar participants under the same conditions (Anvari & Lakens, 2018). Unfortunately, much of what we know in the psychological sciences has either not been replicated or replications have not been attempted, creating an existential crisis (although the exact nature and extent of the crisis invites its own debates; Guttinger, 2020).

This replication crisis in psychology (e.g., Open Science Collaboration, 2015) applies to developmental psychology (Davis-Kean & Ellis, 2019) and to infancy research as well, especially with respect to the small sample sizes of typical studies (Frank et al., 2017). Recent, large-scale, multi-lab efforts have begun to tackle the replication of some key, seminal studies in infancy (e.g., The ManyBabies Consortium, 2020). Because it seems that nonreplicable studies receive more citations than replicable ones (Serra-Garcia & Gneezy, 2021), efforts such as ManyBabies can help the field focus on robust, genuine effects. There is also increased recognition of some reliability problems in infant research (Byers-Heinlein et al., 2021). These researchers argue that observed effect sizes in infancy research are typically low, with direct impact on statistical power. This, in turn, has a negative effect on theory development.

A related problem to replicability is the robustness of findings (Crivello & Poulin-Dubois, 2018). When seminal findings in the field prove difficult to replicate, we need to rethink not just the original study but also the body of work, empirical and theoretical, that ensued. Replicability is certainly desirable, yet it does not imply robustness (e.g., the effect is only observed in a subset of tasks or with a subset of stimuli that should elicit it). And robustness in itself does not tell us that we are measuring what we think we are measuring, because we can easily replicate robust task demands that substitute themselves to the actual phenomenon under study (Sirois, 2022). We need to pay more attention to the task and procedures that we use to understand what (if anything) is revealed by replicable and robust findings (Poulin-Dubois et al., 2018). Increased reliability in measures will help (Byers-Heinlein et al., 2021), especially if this is grounded in theory about the infant mind (Sirois, 2022).

Often overlooked is the contribution of the choice of analyses to the resulting findings. In a rare exploration of this issue, a study crowdsourced 29 teams to analyze a single question from a multivariate dataset about the probability of soccer referees penalizing players more if their skin tone is darker (Silberzahn et al., 2018). Estimated effect sizes of the dataset varied between teams, and whereas 20 teams found a statistically significant effect, 9 did not. In the case of that specific project, there was no onus on teams to return a statistically significant outcome in order to increase publication odds. Clearly, even in the absence of this productivity pressure, decisions about analyses can have a real impact on whether and how findings are disseminated in a research field. However, the researchers identify two potential benefits of multiple analyses of discrete datasets. Firstly, some datasets may naturally invite

Table 1
Various methods of pre-processing and analyzing pupillometry data in a selection of recent infant studies.

Authors*	Pre-processing / filtering**	DV	DV type (primary analysis)	Baseline correction	Primary analysis
Addyman et al. (2014)	Low-pass filter / b-spline smoothing	Peak diameter difference	Discrete	No	ANOVA
Calignano et al. (2023)	Filetering extreme values	Raw diameter	Continuous	Yes	Nonlinear mixed-effect regression
Chen & Westermann (2018)	Low-pass filter	50 ms bin averages	Continuous	Yes	ANOVA (Monte-Carlo)
Csink et al. (2021)	None reported	Raw diameter	Continuous	Yes	Permutation analysis
Fawcett et al. (2017)	Moving average	Average	Discrete	Yes	Mixed-effect regression
Geangu et al. (2011)	Low-pass filter / b-spline smoothing	Spline functions	Continuous	Yes	Functional data analysis
Hellmer et al. (2018)	Moving average / normalizing	Average	Discrete	Yes	t-test
Hepach and Westermann (2016)	Difference filtering / low-pass filter	Average	Discrete	Yes	Generalized linear mixed models
Hochmann & Papeo (2014)	Cluster mass test	Average	Continuous and discrete	Yes	Cluster mass test, ANOVA
Jessen et al. (2016)	Normalizing between conditions	Average	Discrete	No	ANOVA
López Pérez et al. (2020)	Low-pass filter / moving average	Average	Discrete	Yes	ANOVA
Morita et al. (2012)	Savitzky-Golay filter	Time-series	Continuous	Yes	Functional data analysis
Upshaw et al. (2015)	Moving average	Average	Discrete	Yes	t-test
Verschoor et al. (2015)	Low-pass filter / threshold limits	Average	Discrete	Yes	ANOVA
Zhang et al. (2018)	Threshold	Raw diameter	Continuous	Yes	Multiple t-tests using pFDR correction

* excludes papers from the first author
** excludes procedures for the interpolation of missing samples

different analytical strategies more than replications through new observations. This may be most relevant when robust data tend to always be analyzed the same way, study after study. Secondly, convergence and/or divergence of analyses can be uniquely useful when assessing controversial questions or when competing theories make different predictions (especially if those stem from different approaches to data analysis). A more recent iteration of the crowdsourcing approach to discrete dataset analysis found even more variation between the conclusions that could be drawn from different methods (Schweinsberg et al., 2021).

The aim of the current paper was to apply this crowdsourcing approach to pupillometry. For over a decade, pupillometry has grown in popularity within infancy research (Jackson & Sirois, 2022). Pupillometry uses changes in pupil diameter as an index of information processing, an approach particularly useful to study pre-verbal infants. It can be used as a complementary measure to the more common looking time measure in Violation-of-Expectations (VoE) paradigms. In a typical VoE study, infants are familiarized to a stimulus (or set of stimuli) and tested on stimuli that are either familiar or novel relative to those from the familiarization phase. Some measure of interest (typically, relative looking durations) is used to assess whether infants discriminate between types of test stimuli. Unlike looking times, which are a discrete, distal measure of information processing (i.e., how long do infants look *after* some event of interest has happened), pupil diameter is a continuous variable with time-locked properties. If infants react to some event, their pupils dilate in close temporal contiguity with the event of interest. This property is particularly useful to address some general concerns stemming from looking time studies with the VoE paradigm (e.g., Rubio-Fernández, 2019).

The recent popularity of pupillometry is associated with variable approaches to its analysis. For example, Table 1 lists a number of recent pupillometry studies that highlights how preprocessing of raw data, how there are transformed into a dependent variable, whether it is treated as continuous or discrete, and whether it is baseline corrected varies between labs and/or studies. These differences invite different statistical tools for primary analyses of the data. An outstanding question is whether, where applicable, these methods would converge in their conclusions.

In this paper, a study of goal-attribution in infants is used for a collaborative analysis approach. The raw data, collected by the first two authors, was made available to 6 teams who expressed an interest following a call for collaborators on various listservs by the first author. Teams were required to have at least one published paper with their preferred method, and that this method differed from other candidates. In case two or more teams proposed the same method, we planned to use a publication primacy criterion to choose between teams; namely, the team with the earliest published use of the method would be invited. However, we did not need to do so as no two candidate teams proposed the same method. Lead authors submitted the list of potential collaborators and methods to the Editor for approval prior to issuing a formal invitation to participate. A detailed timeline of this process, along with the decisions and procedures put in place and the key information shared publicly (and with teams privately) can be found in the Appendix.

The next section introduces the study used for this collaboration. For this project, we aimed to use a real dataset of infant data from a typical study in the field. We did not wish to use published findings that could a priori inform or constrain analyses. Finally, we needed data that naturally lent themselves to pupillometry analysis. The lead authors had a dataset from a replication of one experiment from Woodward (1998) that satisfied these three constraints and was thus chosen for this project. The next section introduces the selected topic for this collaboration.

1.2. Goal attribution in infants

Human actions are not only a succession of simple actions but also at times a sequence of behaviours organized toward a goal that can be concrete or abstract (Woodward, 2009). Goal attribution has been widely reported in children and infants, leading researchers to wonder if this ability is innate (Csibra, 2010) or acquired (Karmiloff-Smith, 2012). Understanding goal attribution is of the utmost importance in infant development. It is one of the prerequisites for Theory of Mind (ToM: Biro & Leslie, 2007; Csibra & Gergely, 1998; Leslie, 1994). Indeed, 6-month-old infants' ability in a goal attribution task predicts their performance in a ToM task at 4 years of age, even when language level is controlled (Aschersleben et al., 2008).

Research on infants' goal attribution abilities started several years ago using a variety of methods. With an imitation paradigm, Meltzoff (1995) showed that 18-month-old infants can correctly imitate an action done by a human agent, even if it is incomplete, but they do not imitate an incomplete action done by a mechanical device. These results have been replicated using an orangutan puppet (Johnson et al., 2001). Fifteen-month-olds (Johnson et al., 2001) but not 12-month-olds (Bellagamba & Tomasello, 1999) showed the same pattern of results as Meltzoff reported. Seven-month-old infants imitate an adult's action only if the action is not ambiguous (Hamlin et al., 2008). Imitation and action play an important role in the development of goal attribution and in the understanding of others' intentions (Woodward, 2005; Sommerville et al., 2005). Their previous experiences with the same agents can help infants to interpret and anticipate the agents' future behaviours (Moll et al., 2006; Saylor & Ganea, 2007). It has been suggested that goal attribution is a human-specific behaviour (Tomasello et al., 2005; Tomasello & Rakoczy, 2003). Many studies interpret their results in favor of an early ability to attribute goals to social agents within the first year of life.

Research suggests that infants create expectations when they see someone displaying interest toward an object by looking at it, pointing it, or touching it. For example, 12-month-old infants look longer when they see someone looking and smiling at an object and choosing another one instead than when they see someone choosing the object they seemed to be interested in (Phillips et al., 2002). Similarly, Woodward (1998) familiarized 9-month-old infants with a hand grabbing one of two toys. In the test phase, toys' locations are inverted. Infants look longer when the hand grabs the new toy standing on the old location compared to the old toy at the new location. The hand seemingly changing its goal is suggested to be more surprising for infants than the hand changing its trajectory to grab the familiar goal. According to Woodward (1998), infants understand that actions are motivated by goals and attribute a goal to the hand (Woodward, 1998, 2009).

Woodward's (1998) study has been replicated using several variations in order to find the "Woodward-effect": pointing and poking

(Biro& Leslie, 2007; Woodward & Guajardo, 2002), grasping (Woodward, 2003), touching with the back of the hand (Kiraly et al., 2003), lifting (Biro & Leslie, 2007), or only looking at the object of interest, (Johnson et al., 2007). It has also been replicated with non-human agent approaching the goal (Schlottmann & Ray, 2010; Shimizu & Johnson, 2004). Thereafter, the question was what cue do infants need to attribute goals. Subsequent studies tried to find out which characteristics of the scene, agent or goal are the most relevant to attribute a goal to an agent.

Infants under 12-months do not attribute goals if an agent only looks at the goal, but they do if the agent grasps it (Woodward et al., 2001; Woodward, 2003). Goal attribution is often narrowed to some agents (e.g., a familiar person), and does not generalize to other agents (Buresh & Woodward, 2007; Sodian et al., 2004). Nevertheless, goal attribution can be observed with inanimate agents (Csibra, 2008; Csibra et al., 1999; Falck-Ytter et al., 2006; Gergely et al. 1995), or abstract agents (Johnson et al., 2007; Kamewari et al., 2005; Southgate et al., 2008). This ability has been shown as early as six months of age (Luo, 2011; Luo & Baillargeon, 2005).

According to Kiraly and colleagues (2003), 6-month-old infants can attribute goals to unfamiliar actions (displacing an object using the back-of-the-hand), but they need cues of goal-directedness such as equifinal variations of action or salient change of state in the object acted upon. It is similarly argued that infants can attribute a goal to a non-human object if the agent displays the ability to engage in contingent interaction (Deligianni, Senju, Gergely, & Csibra, 2011; Johnson, Shimizu, & Ok, 2007). Furthermore, infants can follow a novel object's gaze if the object has a face or if it reacts contingently with infants' babbling or movements during a familiarization phase (Johnson, Slaughter, & Carey, 1998).

More than the persistent acting on the targeted object or the role of self-propulsion (Luo & Baillargeon, 2005; Luo & Beck, 2010), Hernick and Southgate (2012) examined the important roles of selectivity and efficiency of action. In a one-object situation, if the agent efficiently overcomes obstacles such as making detours (Hernick & Southgate, 2012) or opening a box in order to get the goal (Biro et al., 2011), or if the agent has to modify its path (Csibra, 2005; Luo & Baillargeon, 2005; Luo, 2011), it is considered by infants as goal-directed.

Some researchers showed that producing goal-directed action helps infants to learn and subsequently view other's action as goal-directed. Infant's active engagement plays a role in understanding goal attribution (Gerson & Woodward, 2010, 2014; Sommerville et al., 2005; Sommerville et al., 2008). Researchers also examined that infants do not confound the tool to reach the goal and the goal itself (Sommerville et al., 2008; Woodward, 2005). For example, 26-month-olds have refined goal-processing abilities, as revealed by helping behavior that discriminates agents' intentions when they observe different types of goal interruptions (Green et al., 2021). This ability begins to approximate a causal definition of intentionality ascribed to adults where an agent is believed to have done X because the agent had a stake in the occurrence of X (Quillien & German, 2021). But where does it come from?

For Behne and colleagues, studies using Woodward's paradigm should be interpreted in terms of goal-directed action only if they include a comparison between deliberate and accidental actions (Behne et al., 2005). According to these researchers, studies using habituation paradigms simply demonstrate infants' ability to discriminate two visual scenes. Their own study showed that infants as young as 9 months discriminate between different goals (experimenter unable or unwilling to give an object) leading to the same outcome (the experimenter does not give the object). Carpenter and colleagues (1998) showed that infants from 14 to 18 months of age can distinguish between actions done deliberately or accidentally, and imitate only the deliberate one. Moreover, Reid and colleagues showed that 8-month-old infants are sensitive to unfinished actions (Reid et al., 2007). Goal attribution has also been linked to the mother's interaction style (Hofer et al., 2008), suggesting that social learning is an important part of the way this ability develops (see also Lemche et al., 2007).

An increasing number of experimental studies point out that results from habituation paradigms could be explained by learning mechanisms rather than to complex cognitive abilities (Cohen, 2004; Bogartz et al., 2000; Jackson & Sirois, 2009, 2022; Kagan, 2008), even for social cognition tasks (Perner & Ruffman, 2005; Sirois & Jackson, 2007). Infants could have learned how to behave or to react in different social situations by observing the daily interactions between people. For example, pointing could be seen as a good way to have the adult naming the object or giving it to the infant. Corkum and Moore (1998) showed that gaze-following can be partially conditioned in eight- to nine-month-old infants so they can follow gaze spontaneously.

A methodological problem in many paradigms is that incomplete data are analysed (Sirois & Jackson, 2007; see also Bogartz et al., 1997). Testing all the combinations of variables involved is important. For example, in Wooward's (1998) task, research design and analysis should assess test trials such as the hand grabbing the familiar toy at the familiar location or the new toy at the new location. Some solutions have been proposed. Cannon and Woodward (2012) present the same paradigm, but the hand stops its trajectory just before reaching one toy during the test phase. They compared looking times on each area of interest and showed that infants expect the hand to grab the familiar toy. However, this is at odds with a recent study from multiple labs that show that infants seem to anticipate the paths of actions rather than the goals (Ganglmayer et al., 2019). One experiment in that paper was a direct replication of Cannon and Woodward (2012) but failed to replicate the latter's findings.

Importantly, behavior consistent with goal understanding need not imply goal understanding per se from the onset. Indeed, Gumbsch and colleagues (2021) show that a statistical learning model of goal-directed eye-gaze based on a probabilistic model of dynamical events coupled with outcome prediction naturally produces behavior akin to goal-anticipation without building such a mechanism in the system. Learning processes can bootstrap infants in "goal processing" prior to any conceptual understanding of goal. In other words, witnessing (and learning from) goal-directed behavior can create the anticipation of "goals" in the absence of goal understanding or representation. Yet many studies, despite manipulating perceptual properties of stimuli in ways that would welcome a statistical learning analysis, prefer to ascribe complex cognitive skills to young infants (Geraci et al., 2022), sometimes as early as 5 months of age (Choi et al., 2022; Ting et al., 2021). The issue of where these skills might come from is typically eschewed, at odds with how the brain may well develop these skills ontologically (e.g., Quartz & Sejnowski, 1997). Crucial though is that this debate rests on a coarse dependent measure, looking times.

Measures of cumulative looking time have been highly criticized due to its lack of sensitivity to information processing dynamics (e.g., Haith, 1998; Kagan, 2008; Sirois & Jackson, 2007). It is, in effect, used as a coarse classification tool that identifies which subsets of stimuli infants discriminate. It can be unclear, for example, what processes are at play when infants look at one type of event for 8 s and another type for 10 s *after* the event has ended. In order to improve the method, it has been suggested to use microanalysis (Aslin, 2007) or to use several tools and cross-reference the data (Kagan, 2008). Pupil dilation is a good candidate as an alternative tool for assessment. It has already been used with infants and violation of expectations paradigm (Gredebäck & Melinder, 2011; Jackson & Sirois, 2009, 2022; Sirois & Jackson, 2012). Pupil diameter is also easy to collect while collecting looking times with an eye tracker (as the majority of eye trackers estimate and provide pupil diameter along with gaze data). Pupil diameter increases with cognitive effort and is a time-locked involuntary response (Beatty & Lucero-Wagoner, 2000). Thus, pupil dilation is a dynamic process and allows observing infants' responses in relation with events unfolding in time.

1.3. Overview

The present study replicated Woodward's (1998) paradigm with some methodological changes, and by measuring pupil dilation as well as looking time and gaze position. In Woodward's (1998) original experiments, two distinct toys were placed on a stage next to each other. Infants were habituated to a hand moving to and resting on one specific toy at the same location for several trials. The hand initiated motion from the right side of the display. After habituation was achieved (or after 14 at most habituation trials), they were presented with two test events in which the position of the toys was swapped relative to the habituation phase. In one test event, the hand followed the familiar path to the old location to rest on the novel toy. In the other test event, the hand followed a novel path to the new location to rest on the familiar toy. As outlined, the two key variables in habituation and test events, toy and location, were confounded in the test phase. The novel toy that provided the litmus test of goal attribution was systematically at a familiar location. And the path that the hand must make to reach either location, a dynamic event, may be an important driver of infants' attention (Ganglmayer et al., 2019).

In this replication, we used all 4 combinations of target toy (familiar vs novel) and target location (familiar vs novel) in order to assess the unique contribution of target and location on infant attention, including the possibility of interaction between both factors. Furthermore, as shown in Fig. 1, the initial position of the hand was between and in front of both toys rather than on one side, to avoid any possibility of biasing attention to one toy or location.

Infants were habituated to one video sequence where, between trials, the hand always moved to rest on the same toy at the same location. Selection from the four video sequences that combine target and location for the habituation phase was random for each infant. At test, infants saw all four videos in random order. Fig. 2 illustrates the habituation and test phases. If this task can reveal goal attribution in infants, as suggested by Woodward (1998), looking times should be relatively larger when the hand rests on the new target, regardless of location. If this ability were to interact with location, in which case we predict an ordinal interaction. we would still expect a significant simple effect of target, with longer looking when the new target is selected. If infant behavior is primarily driven by perceptual features of the task, we would predict a disordinal interaction, with potential simple effects for both independent variables (target and location). While the dynamic aspects of hand motion may be important (Ganglmayer et al., 2019), they may elude

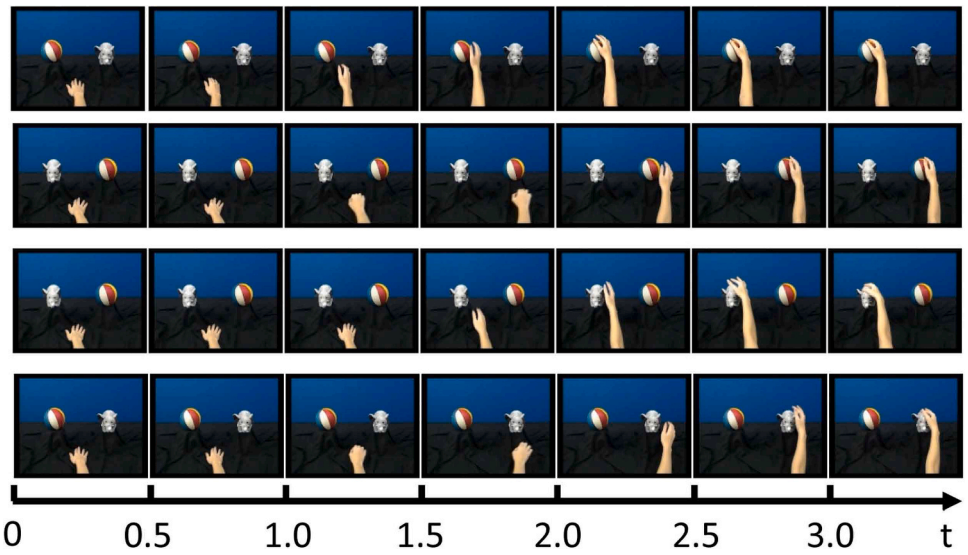

Fig. 1. Frame captures from all 4 stimulus videos from the first three seconds of playback. Each row represents one video, distinguished by target position and target toy. The videos represent Left-Ball, Right-Ball, Left-Elephant, and Right-Elephant respectively. Frame captures are taken at the same 500 ms intervals from the beginning of playback and reveal minor timing differences between sequences. Full video files are available at the OSF repository of this project.

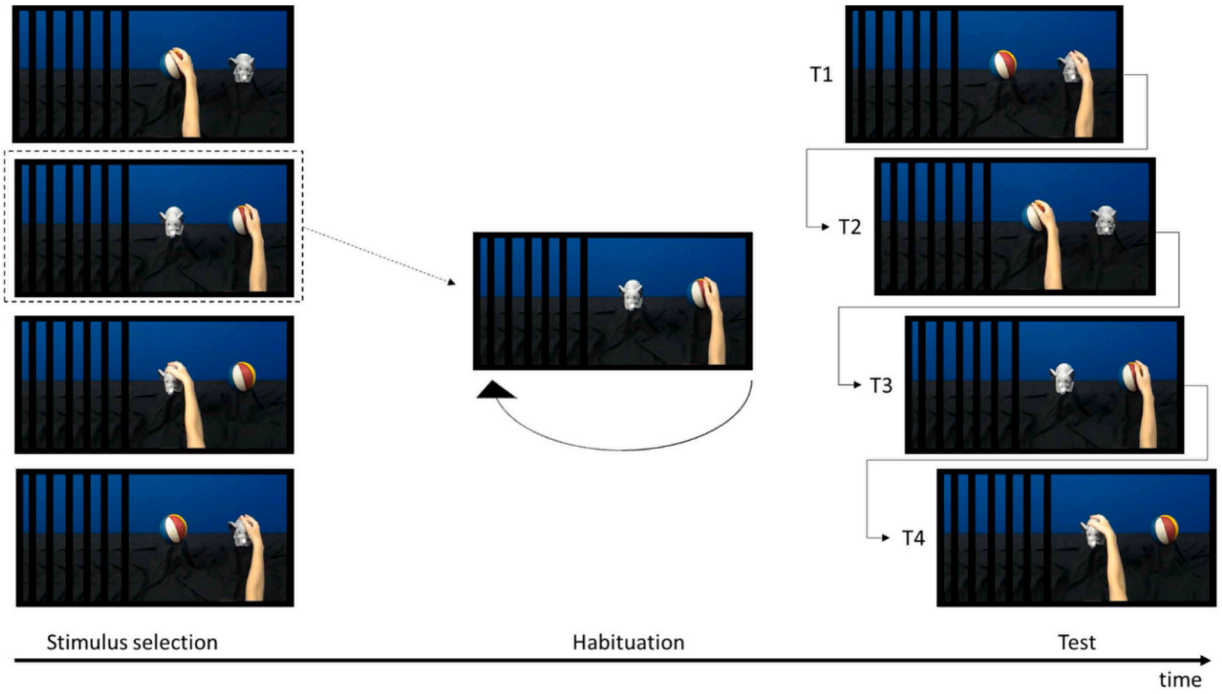

Fig. 2. Schematic illustration of the task. For each infant, one of the four video sequences was randomly selected for habituation. This video was played repeatedly until habituation criterion was met or 20 trials had been completed. At test, infants saw all four videos, one at a time, in random order. In this example, test order is T1: new target – familiar location, T2: familiar target – new location, T3: familiar target – familiar location, and T4: new target – new location. Familiarity and novelty are relative to the habituation video.

the coarse nature of looking times. Based on previous work (e.g., Sirois & Jackson, 2012; Jackson & Sirois, 2022), pupil dilation is expected to provide a finer-grained picture of infants' reactions to apparent changes in goals. Specifically, if the task measures goal attribution, larger pupil diameter is predicted at the time the hand reaches the new target, regardless of location. Again, there could be an ordinal interaction between target and location. However, should perceptual features rather than goals drive infants' information processing, we predict effects for both target and location, and a potential disordinal interaction. Uniquely, because of the temporal nature of pupil diameter and the purported importance of path (Ganglmayer et al., 2019), this could be identified before the hand reaches the target toy, when the path becomes obvious early in the hand's motion. In the context of this special collaboration, an important question is whether different methods converge or diverge in their analysis of the same dataset. In a sense, this project is different from projects such as carried out by teams such as ManyBabies, where multiple labs carry out replications in parallel following an agreed upon standardized protocol (https://manybabies.github.io/about/). This project invites different labs to inspect a unique dataset with their respective and distinct analytical toolboxes in order to compare the outcomes.

2. Method

2.1. Participants

Thirty infants (15 girls, 15 boys) were included in the sample, all of whom were full-term and healthy. Mean age of infants was 282.9 days (SD = 8.10 days). Twenty-two participants were recruited in the greater Trois-Rivières and Shawinigan area (Canada), and 8 participants were recruited in the greater Paris area (France). Participants were recruited through vaccination clinics, parenting classes and advertisement in local papers (Canadian participants), or by letters sent to lists of parents obtained through government's birth lists (French participants). Thirty-two additional infants began the study but were excluded in final analyses due to fussiness or inability to complete the procedure due to crying ($n = 13$), failure to complete test phase ($n = 13$), or technical problems such as tracking loss ($n = 6$).[1]

[1] In Woodward (1998), the overall rate of attrition due to experimental error across all 4 experiments was 0.125, which is lower than 0.2 in this study. However, focusing exclusively on 9-month-olds in Woodward (1998), who can be found in Experiments 1 and 3, the proportions of attrition due to error are 0.22 and 0.31, respectively. Attrition due to infant crying / fussiness is more difficult to compare. Woodward (1998) reports that some infants were sat in seat whereas others in a parent's lap. Our (lead authors) experience testing infants with eye tracking is that the parent's lap increases retention at the expense of data quality. We favor setting up participants in a secure infant seat in front of the apparatus, which allows for better data on average but with increased failure rates in completing tasks due to crying or fussiness.

2.2. Apparatus

The experiment took place in a dimly lit cubicle. Stimuli were presented using E-Prime 2.0. Looking time and pupil diameter data were collected using a Tobii X120 eye tracker (Tobii Technology, Stockholm, Sweden) for the Canadian participants. The eye tracker was positioned beneath a 60 x 34 cm presentation monitor (resolution: 1920 ×1080 px, refresh rate: 60 Hz). Display events filled a 22 cm wide and 17 cm high rectangle (720 ×540 px) in the middle of the screen, while the remaining area was black throughout the experiment. Data for the French participants were collected using a Tobii T120 eye tracker (Tobii Technology, Stockholm, Sweden) equipped with an integrated 34 × 27 cm screen (resolution: 1280 ×1024, refresh rate: 60 Hz). Display events filled a 17.5 cm wide and 14 cm high rectangle (560 ×448 px) in the middle of the screen, while the remaining area was black throughout the experiment. Both eye trackers were located in soundproofed cubicles.

2.3. Stimuli

The stimuli were movies of a hand going towards one of two objects presented (a stuffed elephant and a ball). Four different movies varied the target (elephant or ball) and the location of the target (left or right). In each event, the hand's starting position was at the bottom centre of the display (Fig. 3). This hand location is different from Woodward (1998), whereby the hand was initially to the right of the pair of objects at the beginning of trials. A still picture of the first frame of the video was presented on the screen; the still presentation then persisted until the infant looked at the display (minimum fixation threshold set at 200 ms), at which point the movie of the hand going towards one of two objects began. Hand movement lasted 3000 ms. When the hand reached the object, it rested on it for the remainder of the movie (9000 ms). At the end of the movie, the next trial began.

2.4. Procedure

Parents and infants were welcomed and given time to adapt to the lab environment. When infants were ready, they were positioned in the testing cubicle, sitting on their parent's lap, within tracking distance of the eye tracker (approximately 60 cm). Five-point calibration was performed, and the experiment began.

During the habituation phase, infants were shown the same movie several times. Mean looking time was calculated from the first three trials, and half of this mean looking time served as habituation criterion. The movie was presented until mean looking time on the currently last three consecutive trials (beginning from the sixth trial) was equal to or less than the criterion (with a maximum of 20 trials). When criterion was met (or 20 trials had been shown), four test trials were presented. Order of the test events was randomised across infants. The experiment ended and parents were debriefed. If infants became fussy or cried during the experiment and stopped looking at stimulus events, the experiment was terminated and parents were debriefed. A video camera above the presentation monitor allowed the experimenter to unobtrusively observe infants throughout the test. All experimental equipment was operated from outside the testing cubicle.

Infants were habituated to one of four possible movies and when habituation criterion was met (or 20 trials completed) they were presented the four test trials. The latter comprised a 2 × 2 (path x target) factorial design. Order of presentation was randomised across infants, and the habituation video was randomly selected. Gaze and pupil diameter data were recorded at a sampling rate of 60 Hz, and the eye tracker was controlled from E-Prime 2.0 (Psychology Software Tools, Pittsburgh, PA). Looking times were computed in E-Prime to control task flow (e.g., assessing habituation) and are the values used in the results section. Raw data from the eye tracker for all trials were saved to a file for each participant and were used for gaze and pupil analyses.

3. Results and Discussion: Looking time and gaze data

Looking time was computed for each participant on each trial. Besides looking time at the entire scene, 2 Areas of Interest (AOIs) were defined as left AOI (200 ×200 pixels corresponding to the location of the left toy) and right AOI (200 ×200 pixels corresponding to the location of the right toy), and looking times on these two target AOIs were also analysed.

3.1. Habituation trials

The mean number of habituation trials was 11.6 (SD = 5.44, range = 7–20). There is no difference between the 4 habituation videos ($F(3,26) = 0.735$, $p = .543$). Mean looking times (sec) at the first three habituation trials was compared to the last three habituation trials (Fig. 4).

A repeated-measures ANOVA was performed with one intra-subject IV (bloc: first or last habituation trial) and two inter-subjects IVs (2 targets x 2 paths). There is no triple nor double interactions between variables on looking time to the whole scene (all $Fs(1,21) < 2.81$, ps $> .109$). There is no main effect of trajectory or target ($Fs(1,21) < 0.39$, $ps > .539$), but infants look more at the first habituation trials than the last ones ($F(1,21) = 37.77$, $p < .001$, $\eta^2_p = .643$).

3.2. Test trials

Mean looking times to the four test trials (familiar target/familiar path, familiar target/new path, new target/familiar path, and new target/new path) are shown in Fig. 5. A repeated-measures ANOVA was performed on mean looking time. There is no interaction

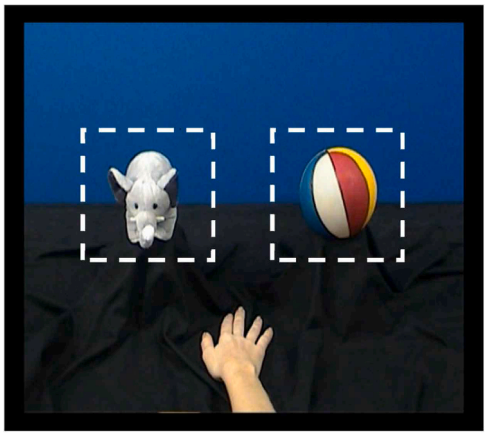

Fig. 3. Still frame from one of the stimulus videos, showing the hand at the start position, below and between the two toys (ball on the right and elephant on the left). The two areas of interest are shown by dashed lines.

or main effect of the looking time on the entire scene (*Fs*(1,24) < 1.46, *ps* > .239). Results do not change if the five infants that have seen 20 habituation trials are removed from sample (familiar target/familiar path: mean = 5.5, SD = 3.41, familiar target/new path: mean = 5.3, SD = 3.48, new target/familiar path: mean = 5.7, SD = 3.89, and new target/new path: mean = 5.8, SD = 3.63). There is no interaction between path and target, nor main effect (*Fs*(1,19) < 19.00, *ps* > .500).

In order to examine looking times in more detail, a ratio was calculated by dividing the mean looking time on each AOI by the mean looking time on the scene. Then data were recoded depending on the familiarization side. Indeed, whatever the path infants were familiarised to (left or right), data can be summarized by saying that in some test trials the AOI of interest (meaning the target the hand goes to) corresponds to the AOI of interest in the familiarization phase (familiar location) and in other trials the AOI of interest is different from the AOI of interest during the familiarization phase (new location). Table 2 shows the looking time ratio for each location (familiar and new) depending on target (familiar or new) and on path (familiar or new).

No interaction was found between target and path or between location and target (*Fs*(1,24) < 1.05, *ps* > .315) but an interaction was found between location and path ($F(1,24) = 8.99$, $p = .006$, $\eta^2p = .273$). When the path is familiar, infants look more at the

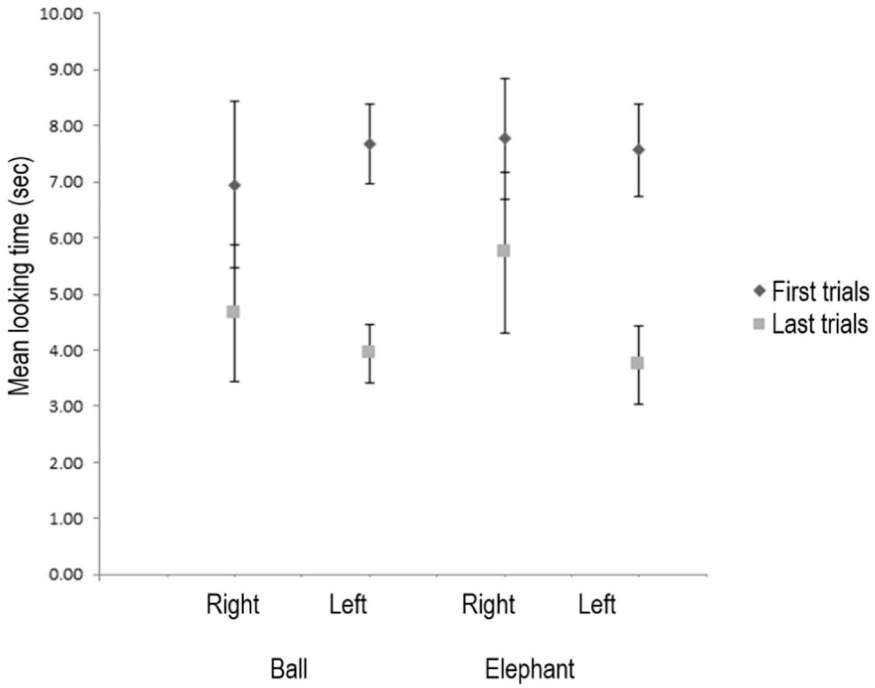

Fig. 4. Mean looking times for first three and last three habituation trials for each condition of habituation (2 targets x 2 paths). Error bars represent the standard error of the mean.

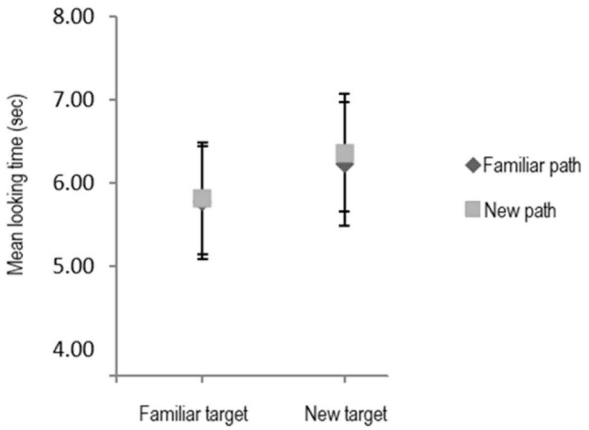

Fig. 5. Mean looking times to the different test trials on the entire scene AOI depending on target (familiar/new) and path (familiar/new). Errors bars show the standard error of the mean.

familiar location than the new one ($F(1,24) = 6.67$, $p = .016$, $\eta^2p = .217$) whereas there is no particular interest in one location when the path is new ($F(1,24) = 2.95$, $p = .099$, $\eta^2p = .110$).

Importantly, the design of this study differs from the original study by Woodward (1998). In the original study, there were only two test events (new target with familiar path vs familiar target with new path). These were presented three times each and looking times were summed for each type of test event. In the current study, we treated target and path as factors and presented the four combinations of these (once each) in a repeated-measures design. This created the possibility that looking times would decrease over testing, reducing the sensitivity of the measure to experimental manipulation. It also created the possibility that the first test trial would generate most interest for infants who saw the new path or the new target (or both) at T1. However, looking times between test trials T1 through T4 (with means (SD) of 5.43 (3.52), 6.89 (3.95), 5.37 (3.36), and 5.36 (3.48) respectively) did not differ significantly ($F(3, 75) = 2.53$, $p = .064$, $\eta^2p = .092$). Likewise, treating the first test trial T1 as a between-subjects design did not reveal any interaction between target and path, nor any main effect of target or path (all $Fs(1,24) < 0.79$, $ps > .384$, $\eta^2ps < .033$). On the basis of these last two analyses, the modification of the design does not appear to have negatively affected the sensitivity of looking times.

3.3. Horizontal gaze

Fig. 6 shows the mean gaze position on the horizontal axis as a function of time (during the movement of the hand- 3000 ms), habituation hand path (left-right) and test trials hand path (left-right). At the beginning of the video, infants from all groups look at the center of the screen (hand location). During the first 1000 ms, they tend to look at the location they were familiarised with, particularly when the hand goes to the opposite side compared to familiarisation side. Around 1000 ms, they follow the hand location and look at the object grabbed by the hand. This last pattern is clearly defined when habituation and test path are congruent and slightly attenuated when habituation and path are incongruent.

3.4. Discussion

Concerning looking times, no interaction was found between path and target. When analyses are performed on the entire test trial set (as in this 2 ×2 design), results do not replicate Woodward's (1998) findings. However, using finer-grained analysis of gaze data, it seems that infants anticipated slightly the location the hand will go depending on the habituation side and then followed the moving hand. When the path is familiar, infants look more at the location they have been habituated to than the other, no matter the target. When the path is new, they do not show clear preference for the new location at first but look at the location where the hand goes, as if they expect the hand to go to the usual AOI and then look at the new one when they notice the change. Ratio data reported in Table 2 are thus consistent, at task-demand level, with Woodward's (1998) results. Because the identity of the target is irrelevant, which our factorial analysis reveals, we cannot however infer goal attribution. As showed by Yu, Yurovsky and Xu (2012), visual data mining brings complementary information to traditional statistical analysis. It should be noted that the initial position of the hand in our study

Table 2
Looking time ratio for each test trial depending on the side of the AOI (familiar or new).

	Familiar location		New location	
	Familiar target	New target	Familiar target	New target
Familiar path	.35 (.289)	.26 (.232)	.15 (.185)	.25 (.244)
New path	.18 (.207)	.22 (.245)	.33 (.303)	.28 (.267)

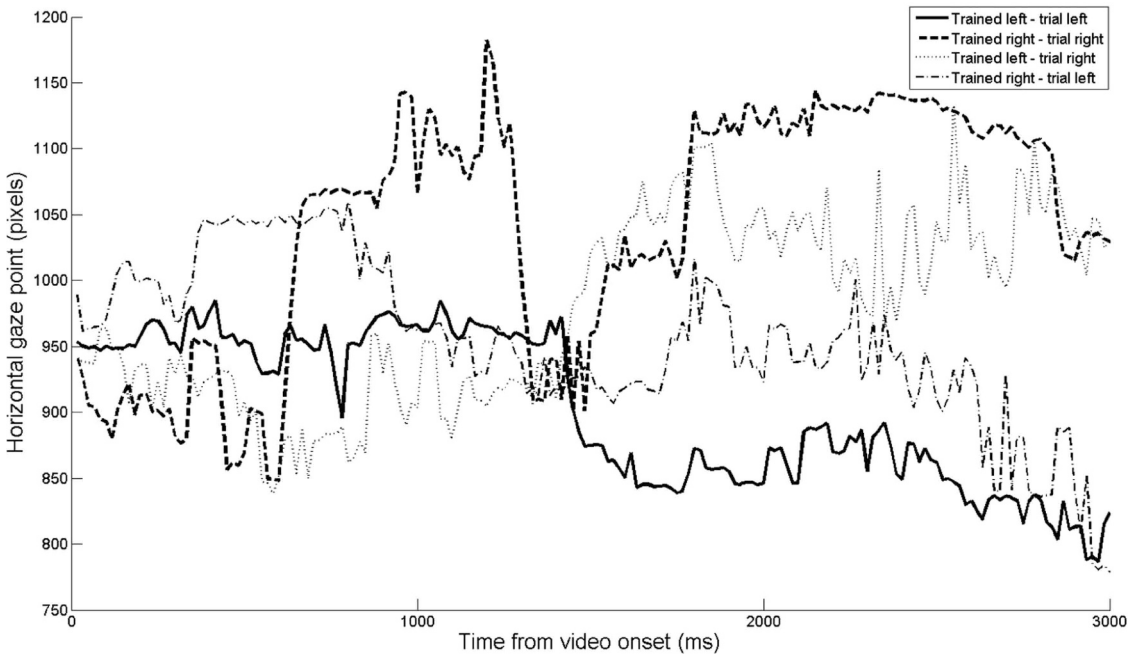

Fig. 6. Horizontal gaze point (values below 960 are towards the left, values above 960 towards the right) as a function of habituation side (left-right) and test trial side of hand grabbing (left-right) during test trials.

was between both targets, rather than to one side of the display. This was used to prevent a position bias, and is a methodological departure from Woodward's (1998) procedure.

4. Results and discussion (Windowed Median Linear Mixed Model - Team A: Blaser, Kaldy, and Donenfeld)

4.1. Pupillometry method

Preprocessing of pupil data was conducted using a custom MATLAB (2021b) script, with subsequent hypothesis testing carried out in GraphPad Prism (9.0). Pupil preprocessing consisted of the following steps:

1) Within the pupil traces from each trial, we defined two phases of interest. The *baseline phase* was the 500 ms immediately preceding movie onset, and the *critical phase*, from which task-evoked pupil responses were measured, was the 2000 ms window starting 1000 ms after movie onset. This 1000 ms offset was selected to give sufficient time for the participant, after movie onset, to perceptually and cognitively assess the path and target, and to allow for the emergence of any subsequent task-evoked pupil response. In the absence of prior data on pupil responses in this task, we could not predict precisely when such a response might occur, so we chose a wide, 2000 ms window, which spanned the remainder of the animated portion of the movie, to minimize the risk of missing a transitory response.
2) Each pupil trace was then regularized to ensure that samples (using EPrime RTTime) were synchronized to 60 Hz. (Data were nominally collected at 60 Hz, but samples can be missing, repeated, or out of sync; this step corrects for these (rare) occurrences.)
3) Then, where (and only where) pupil samples were present from both the left and right eyes, they were averaged to create a single pupil diameter estimate, otherwise the sample was left empty.
4) To eliminate potential spurious values, we then removed outliers within each trace (values >3 Median Absolute Deviations (MAD) from the median; Leys et al., 2013), within a moving window set to 220 ms, nominally the fastest meaningful pupil response (Mathôt et al., 2015).
5) Traces were then screened to ensure there was sufficient data for valid subsequent analyses. In total, 30 participants had been each presented with four test trials, for a total of 120 possible trial traces. From these, 10 traces showed no data for the entire trial. Of the remaining 110 traces, 23 had no pupil data at all in the baseline and/or critical phase. This left 87 traces that had data that could potentially be analyzed. To these, we applied a data quality exclusion, rejecting traces that had a gap of missing data greater than half of the relevant phase (i.e., rejecting traces with gaps >250 ms in the baseline phase and/or gaps >1000 ms in the critical phase). This screening removed 13 traces, resulting in a final set of 74 traces, from 23 participants (down from 30, as 7 participants did not provide valid traces for any of the four possible test trials).
6) For each trace, we then used linear interpolation to fill missing values, followed by gaussian smoothing (with a moving window set to 220 ms).

7) Finally, the traces were baseline-corrected by taking the median pupil value within the baseline phase and subtracting this value from the trace (Mathôt et al., 2018).

4.2. Results

Pre-processing completed, we then determined the *critical pupil value* - the median pupil diameter value within the critical phase of each trial's trace. These critical pupil values were entered into a 2 × 2 repeated-measures mixed-effects model analysis (using a compound symmetry covariance matrix and REML fitting) with *target familiarity* (familiar versus new) and *path familiarity* (familiar versus new) as fixed factors, and *participant* as a random factor. Supporting the validity of this test: 1) the overall set of critical pupil values did not deviate significantly from normality (as determined by a Lilliefors test $p = 0.47$ (Lilliefors, 1967)), 2) repeated-measures matching was deemed effective (chi-square test: $\chi^2(1) = 10.6, p = 0.001$), and 3) the QQ plot did not reveal any substantive deviations from normality of the residuals.

The mixed-effects analysis showed no significant main effect of *path familiarity* ($F(1, 22) = 0.169, p = 0.685$), *target familiarity* ($F(1, 22) = 0.162, p=0.691$), or their interaction ($F(1, 4) = 0.618, p=0.476$). Individual and mean traces are shown in Fig. 7 for each of the four conditions, along with a scatter bar showing the critical pupil diameter values that were used in the mixed-effects analysis (from left to right, top to bottom: familiar target/familiar path ($M = -0.016$ mm, SD = 0.266); familiar target/new path ($M = 0.064$ mm, $SD = 0.354$); new target/familiar path ($M = 0.038$ mm, $SD = 0.215$); new target/new path ($M = 0.008$ mm, $SD = 0.269$)). While the mean pupil diameter was numerically higher when a new path was introduced (0.064 mm versus −0.016 mm, yielding a difference of

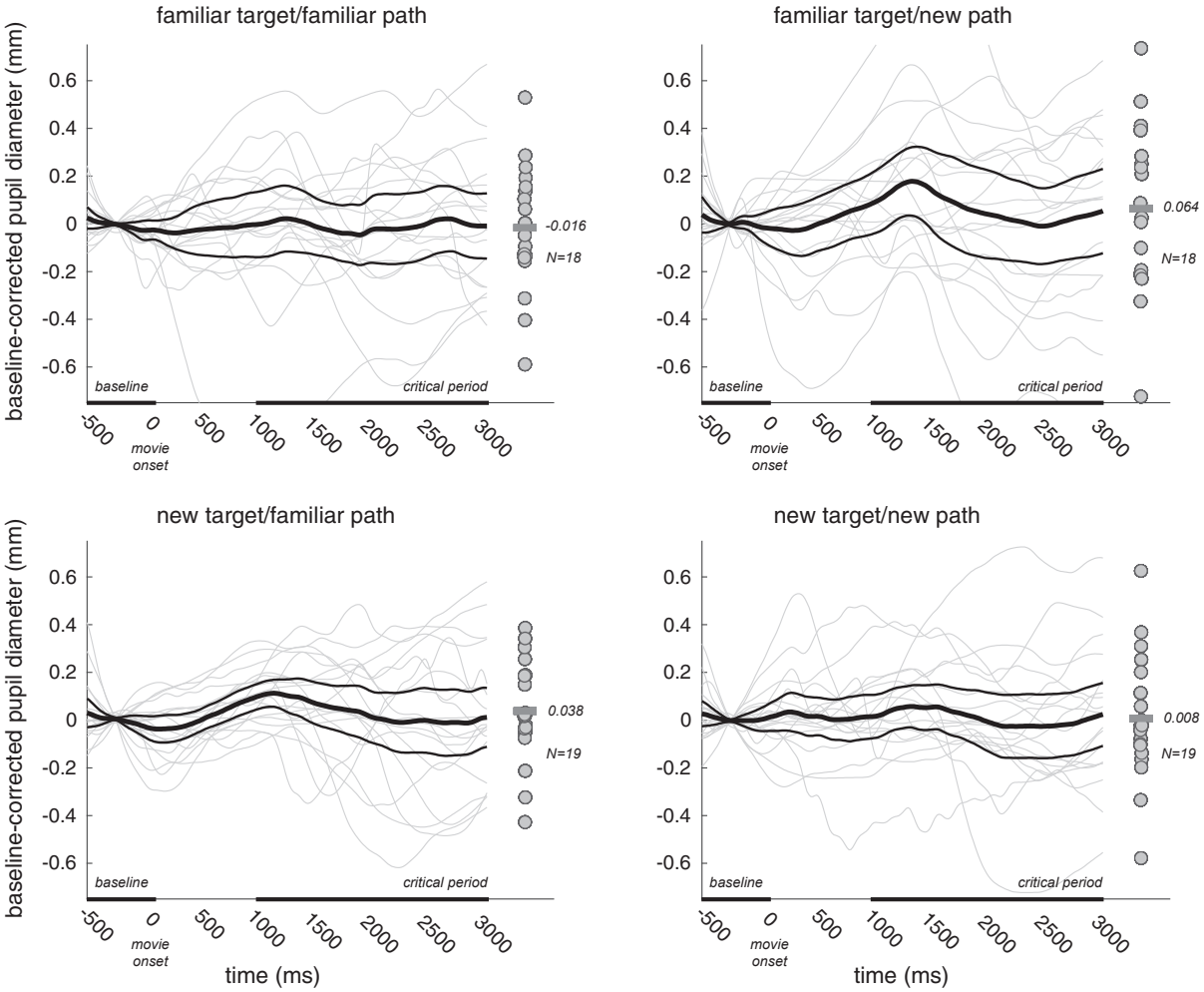

Fig. 7. Baseline-corrected, task-evoked pupil responses. For each of the four conditions, individual traces (light gray lines), as well as mean traces (black lines, with 95% CI bands), are shown. Here, traces are shown beginning at the start of the 500 ms *baseline period*, followed by movie onset (at time point 0). 1000 ms after movie onset, the 2000 ms duration critical phase begins (indicated by the vertical line), during which the median pupil diameter is determined. These *critical period values* are shown in a scatter bar to the right of each panel (gray dots) along with their mean (gray horizontal bar).

0.081 mm and an effect size (Cohen's *d*) of 0.26), this was not significant. Exploratory post-hoc contrasts comparing each of the three new-outcome conditions to the familiar/familiar control (adjusted for multiple comparisons using Dunnett's method) did not reveal any significant effects (familiar target/familiar path versus familiar target/new path, Dunnett's $q(4) = 0.831$, $p = 0.754$; versus new target/familiar path, $q(4) = 0.273$, $p = 0.984$; and versus new target/new path, $q(4) = 0.011$, $p > 0.999$).

4.3. Discussion

In sum, our analysis did not find any significant differences between task-evoked pupil responses in the four test conditions. Neither a new path, a new target object, nor their combination evoked a significantly greater pupil response relative to the condition with familiar outcomes. That said, while retrospective power analyses are challenging (Hoenig & Heisey, 2001), since there were only 18–19 data points for each condition, the present data set is likely underpowered for strong conclusions. As well, the present analyses should be considered quite conservative, as they are based only on a single summary pupil value, removing important information about change over time that would be captured by FDA analysis (Jackson & Sirois, 2009); only relatively large, sustained effects will be detected. In our previous work, we have used methods based both on summary values (Blaser et al., 2014) and FDA analyses (Cheng et al., 2019), and found that the latter better utilizes the advantage of the task-evoked pupil response as a measure: it reflects real-time modulations of cognitive effort.

5. Results and discussion (Multiverse - Team B: Calignano, Russo, and Valenza)

5.1. Pupillometry method

Pupillometry, like most of psychophysiological and behavioral measures, implies a number of degrees of freedom in data processing before statistical data analysis (Simmons, Nelson, & Simonsohn, 2011, Wicherts et al., 2016). The multiverse approach aims at dealing with the uncertainty introduced by such data management to evaluate the robustness of statistical results across a set of plausible alternative preprocessing choices (Steegen et al., 2016). In the multiverse approach, data preprocessing is seen as a *garden of forking paths* (Gelman & Loken, 2014), where each processing step implies a decision that might determinate the subsequent statistical results. In other words, rather than being significance-oriented, the leading question of the multiverse approach is whether the estimated results are robust or driven by specific decision made in data preprocessing (Dragicevic et al., 2019).

In the present analysis, we dealt with three plausible *forking paths* characterizing pupil data preprocessing, namely (1) the filtering of pupil size datapoints, (2) the selection of areas of interest differently implemented across the two labs (i.e. France and Canada) and (3) the baseline correction (Calignano, Girardi, & Altoé, 2023). By comparing the results obtained from alternative universes of data, we aimed at determining the most robust patterns of findings while weighting the impact of preprocessing choices in driving the results.

5.1.1. First degree of freedom: Pupil size filtering

We calculated the average of raw pupil diameter values from the two eyes when the eye tracker got a good signal from both eyes (validity = 0). Otherwise, measurements where only one eye was recorded with a good signal data were excluded.

Fig. 8 shows a basic scatter plot depicting pupil size of the two eyes as acquired during the whole data collection. Given that cut-off values are usually applied according to average human physiology (e.g., Mathôt et al.; 2018), we moved a first step into the multiverse of data processing by building an alternative dataset which only included pupil size values higher than 2 millimeters (step 1, choice A: filtered data), while at the same time keeping the full dataset into consideration (step 1, choice B: unfiltered data). This first step allowed to check to what extent the variability introduced by extremely small yet plausible positive values (<2 millimeters) might drive the results interpretation at the trial and the subject level (Mathôt et al.; 2018).

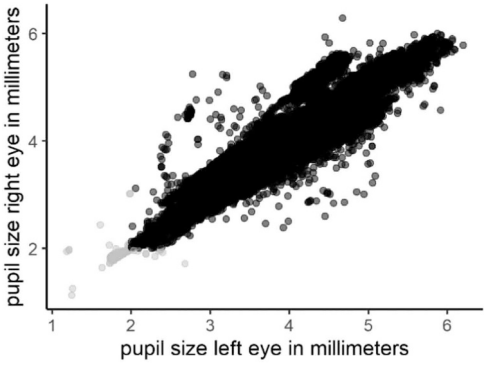

Fig. 8. Scatter plot correlating left and right eye's pupil size in millimeters. Gray points indicate the values excluded in the second filtered dataset.

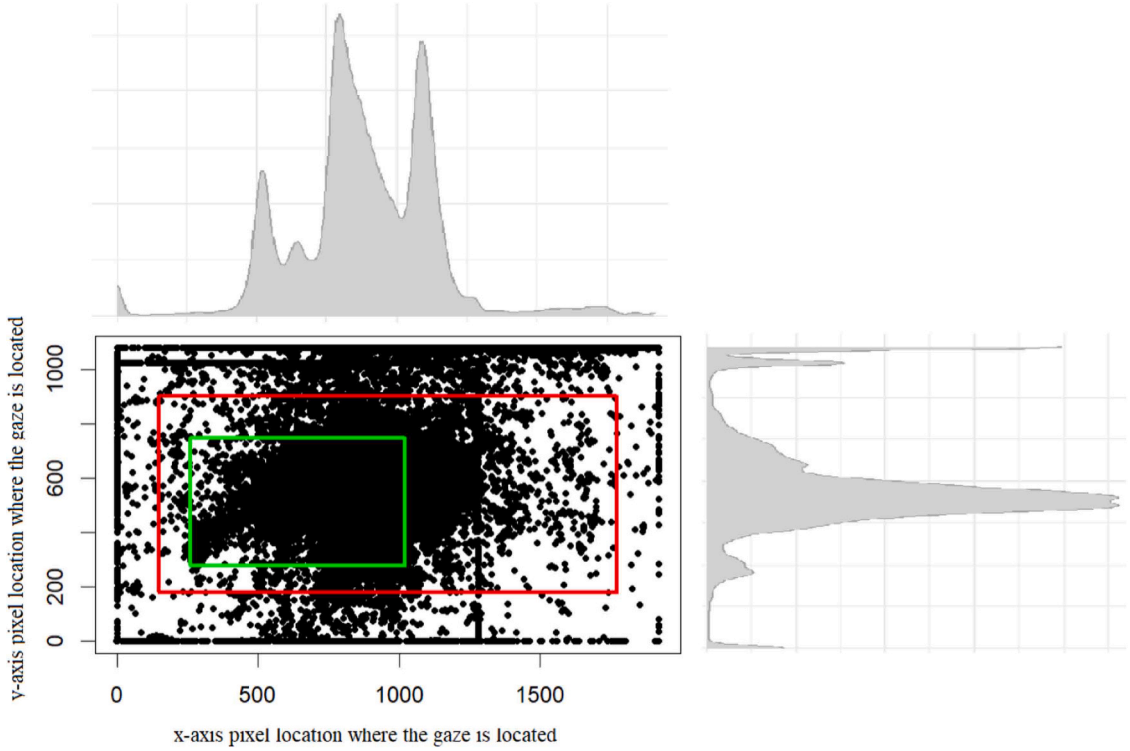

Fig. 9. Gaze-points coordinates corresponding to the X and Y axis' pixel location, where the gaze is located. Colored rectangles indicate the area of interest used by the Canadian (red) and the French (green) lab, respectively. Density plots of the Gaze-points X and Y in arbitrary units are plotted in gray.

5.1.2. Degree of freedom 2: Area of interest

Fig. 9 shows all gaze data mapped into a 2D coordinate system corresponding to the whole eye-tracked space resulting from the two data collections implemented in the French and the Canadian lab, respectively. It can be noted that the data points are distributed within two frames corresponding to the resolutions used by the two labs (i.e., larger resolution in the Canadian than in the French laboratory setting). Coherently, the size of the area of interest (AoI) differed between the two labs.

In the previous step into the multiverse, we obtained two datasets (filtered and unfiltered data) starting from the whole data collection. Here, we moved a second step deeper by focusing on gaze data coordinates. That is, we alternatively subset the data within the AoI implemented in the French (step 2, choice A: French AoI) and the Canadian laboratory setting (step 2, choice B: Canadian AoI), respectively. In doing so, we added a second *forking path* to the multiverse analysis so obtaining four datasets (i.e., step 1 $\times$ step 2).

5.1.3. Degree of freedom 3: Baseline correction

As a final step into the present multiverse, we included a last *forking path* related to the baseline correction. Starting by the four datasets obtained from the previous steps of data preprocessing, we applied two alternative subtractive baseline corrections at the trial level (Mathôt et al., 2018). Plausible values for baseline length were constrained by the fact that, in the present study, each trial begin with a still picture of the first frame of the video that persisted until the infant looked at the display form a minimum fixation threshold set at 200 ms, at which point the movie of the hand going towards one of two objects began. According to the experimental procedure, the median of the first 200 ms was the maximum time window useful to perform baseline data correction. Thus, for each trial within each participant, we corrected the pupil signal by subtracting a baseline segment of either 100 (step 3, choice A: 100-ms baseline) or 200 ms after the stimulus onset (step 3, choice B: 200-ms baseline), in addition to considering the uncorrected signal (step 3, choice C: no correction). Doing so, we obtained 12 plausible datasets from the whole data collection (i.e., step 1 $\times$ step 2 $\times$ step 3).

As an illustrative example, Fig. 10 shows the grand average of pupil size by the goal condition in the case of filtered pupil size values (step 1: choice A) and the Canadian AOI (step 1: choice B), highlighting the differences across the three alternative baseline corrections (step 3).

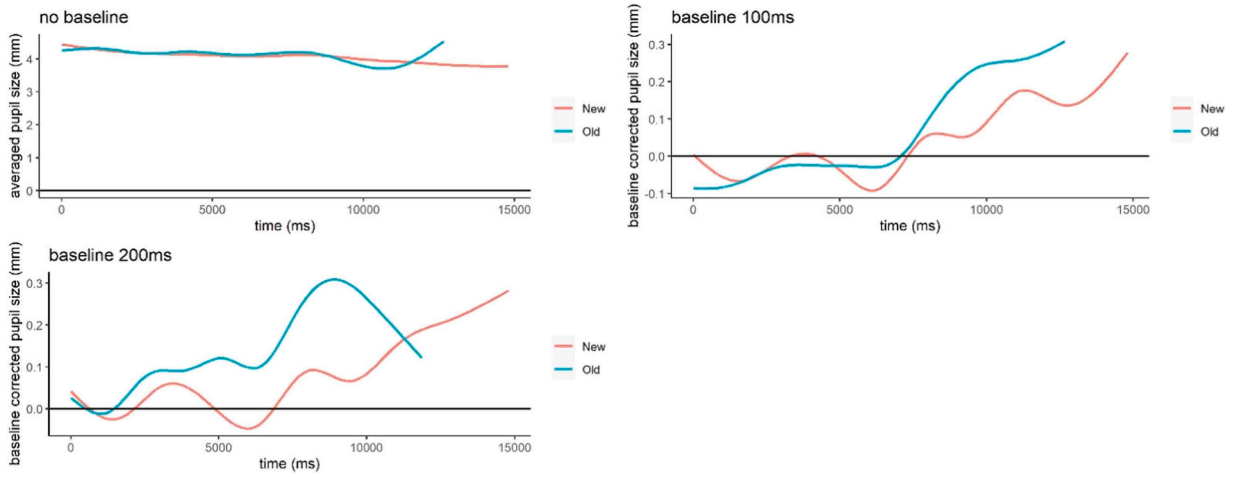

Fig. 10. Averaged pupil size variation (no baseline) and pupil changes relative to baseline (100 and 200 ms) smoothed across time. The red and green lines represent the Old and New target, respectively.

5.2. Results

Fig. 11 shows the number of included data by participants depending on both data filtering (step 1) and area of interest (step 2), corresponding to the total number of raw data included in the statistical analysis by participants. Such sanity check allows to acknowledge the impact of unbalanced and subsample of data in driving the statistical results.

Data were analyzed with R (R Core Team, 2020) using linear mixed-effects models with the *lme4* package (version 1.1–30; Bates, 2010), and *ggplot2* (version 3.3.6; Wickham, Chang, & Wickham, 2016) and *sjPlot* (version 2.8.11; Lüdecke & Lüdecke,2015) to visualize data and results.

We selected the mixed-effects approach that allows (i) to model preprocessed pupil size data by both fixed (i.e., target, path, and time) and random effects (i.e., individual variability), (ii) to fit with both discrete and continuous factors (i.e. time in milliseconds), and (iii) to deal with unbalanced datasets by stabilizing the estimation of the parameters under investigation (Bates, 2010). We modeled the pupil dilation across the 12 datasets by specifying a linear mixed-effect model with a 2 (Target) × 2 (Path) factorial design as a function of time (treated as a continuous predictor). Of note, the Supplementary Multiverse materials show all the plots of the estimated interactions and main effects resulting by the modelling of the whole multiverse of datasets (N = 12) with all the coding materials to fully reproduce the present analysis.

Fig. 12A shows the distribution *p* values of the main and interactive effects obtained by modelling the 12 datasets, and Fig. 12B shows the interaction coefficients estimated during the test phase. It can be seen that the statistical results were overall characterized by a strong robustness to data preprocessing choices.

Indeed, only two models from the dataset with unfiltered data (step 1: A), filtered only for the French AoI (step 2:A) and with no

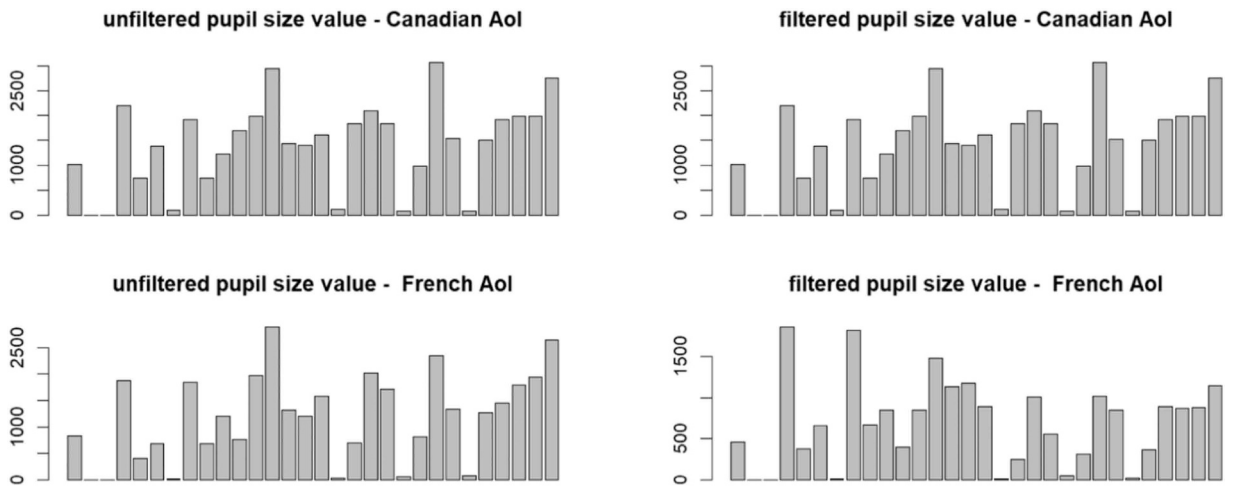

Fig. 11. Number of raw data and participants included in the statistical analysis obtained by alternatively applying the step 1 and 2 preprocessing choices. AoI = Area of Interest).

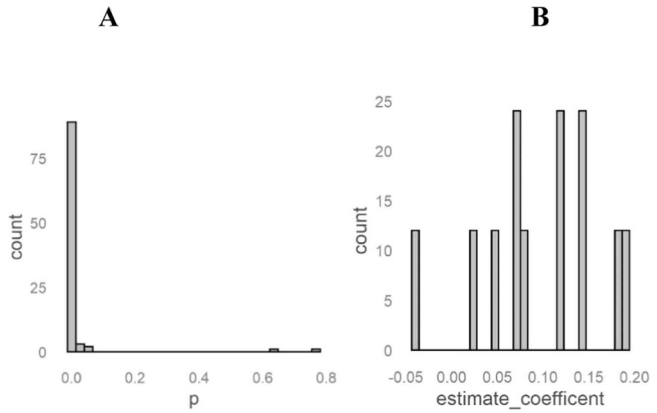

Fig. 12. Histograms of the *p* values of the Target x Path interaction on pupil dilation for the multiverse of 12 data sets (left panel) and the estimated coefficients indicating the impact of the Old vs New goal grabbed via the Same path in predicting pupil size increasing (right panel).

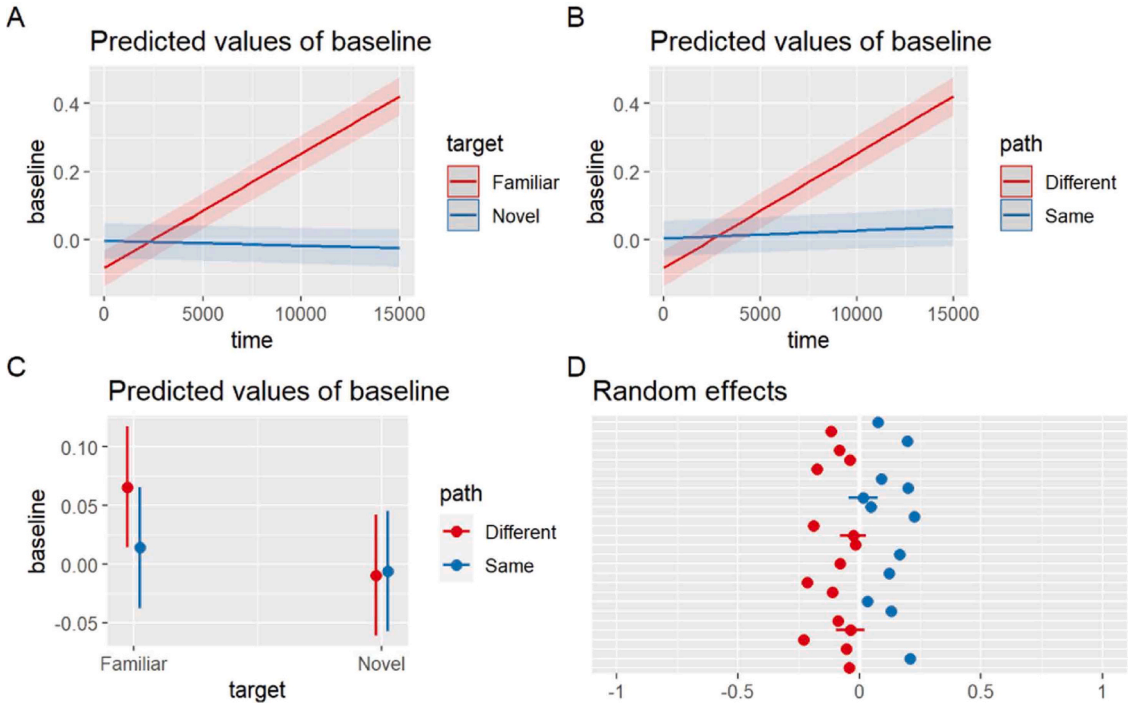

Fig. 13. Effect plot of showing (A) the interaction between time and Target (Familiar and Novel), (B) the interaction between time and Path (same and different), (C) the interaction between Target and Path, (D) the random effects at the individual level, as estimated by the linear mixed-effect modelling the dataset with filtered values, the French apparatus and the 100 ms baseline (i.e., one representative case among the 12 considered universes).

baseline (step 3:A) produce *p* values > 0.05. Nevertheless, across the whole multiverse we found that the Familiar target significantly and consistently predicted higher pupil dilation compared with the Novel target. Moreover, the trials in which the Familiar target was grabbed via a Different path predicted higher level of pupil dilation compared to all other combinations in more than half of the multiverse space, as shown in Fig. 12B and Fig. 13C. As an illustrative example, Fig. 13 shows all interactions and the random effects estimated by the dataset with filtered values (step 1), the French AoI (step 2) and the 100 ms baseline (step 3).

5.3. Discussion

First, the multiverse analysis investigated the robustness of the interaction under scrutiny (Target × Path) showing that infants likely allocate more cognitive resources when the hand grabs the old goal via a new path compared to both a new and the old goal grabbed via the old path, and a new goal grabbed via the new path. The hand changing its path but not its goal emerged to be more

surprising for infants than the hand changing its path to grab a new goal. In other words, the multiverse of results suggests that the hand changing its goal might be less surprising for infants than the hand changing its trajectory to grab the familiar goal. Those patterns of results are openly in contrast with classical findings showing that the hand that grabs the new goal standing on the old location increased looking times in young infants, compared to the old goal at the new location (Woodward, 1998, 2009). However, the present multiverse of results seems to corroborate other findings suggesting that infants expect the hand to grab the old goal (Cannon & Woodward, 2012).

Second, the present multiverse analysis dealt with the uncertainty and noise increased or reduced by specific preprocessing steps. However, we found that the main effect of the familiar goal in predicting higher level of resource allocation (higher pupil dilation) was consistently found in interaction with time course of the test trial, independent from the preprocessing steps. Specifically, the first degree of freedom (step 1: filtered vs unfiltered data), a preliminary robustness check in statistical analysis (for a debate see Reiss et al.; 1997), showed that extremely low positive values (< 2 millimeters) introduced substantial variability yet do not remove the impact of the familiar goal in predicting higher level of resource allocation compared with the new goal. Of note, such first degree of freedom adds knowledge on the robustness of the goals anticipation effect while dealing with physiological variability and possibly, measurement error. Similarly, the second degree of freedom (step 2: Canadian vs French AoI) allowed to extend the familiar goal effect to different lab settings increasing the generalizability of the effect.

The last degree of freedom (step 3: 100-ms vs 200-ms baseline correction vs no correction) related to the baseline correction that implies that pupil sizes are firstly compared with those recoded over a baseline window, nested by trial and participant. Thus, the dependent variable becomes the change in pupil size relative to the median baseline value. Such an approach allows for a within-trial analysis, that is, an analysis in which each trial (nested by subject) is taken into account and considered as a random effect. It is worth noticing that the baseline correction (third degree of freedom) improved the model fit and reduced the uncertainty associated with the effect of interest compared with no baseline correction across the whole multiverse (see Supplementary Multiverse materials).

The present exploration indicated that infants likely anticipate the familiar goal by increasing their resource allocation towards the old (vs new) goal. However, the overall multiverse of datasets shows a high level of heterogeneity mainly due to individual variability (i.e. random effects) and unbalanced datasets, which are the norm rather than the exception in infancy research. It remains open the possibility that the familiar goal effect might be magnified in a different (vs same) location indicating that infants might dynamically shift strategy by anticipating path of actions instead of uniquely relying on the understanding of others actions' goal (as indicated by the model step 1:B × step 2:B × step 3:C, in the Supplementary Multiverse materials). Finally, the multiverse approach aimed at increasing the robustness of the statistical results interpretation by weighting the impact of preprocessing choices on the effect under discussion. Instead of fearing uncertainty inherent in pupillometry, the present multiverse offers methodological and theoretical advices on how to embrace uncertainty while checking the robustness of the statistical results.

6. Results and Discussion (Quantile cut-off filter / Linear Mixed Modelling - Team C: Hepach)

6.1. Pupillometry Method

The data were processed with *R* (version 4.2.0; R Core Team, 2022) and within *R* we used the packages *tidyverse* (version 1.3.1; Wickham et al., 2019), *ggthemes* (version 4.2.4; Arnold, 2021), *ggpubr* (version 0.4.0; Kassambara, 2020), and *RcppRoll* (version 0.3.3; Ushey, 2018). Statistical analyses were conducted with the package *lme4* (Bates et al., 2015). The majority of individual steps for pre-processing were based on previously published routines (Hepach et al., 2012, 2020) and the decision about which post-processing analysis time windows to set for the baseline- and test-phase were based on a recent review of pupillometry in developmental psychology (Hepach, forthcoming) The details of pre-processing (filtering the raw data) and the post-processing (calculating the dependent measure per subject per trial) are provided in the accompanying *R*-script (*pupil_procs_Step1-Pupil_IBAD.R*) at the OSF repository for this project.

6.1.1. Pre-processing

For each subject, the raw data were imported into *R* and only the columns relevant to the subsequent analyses were included. For the two columns relating to the pupil size in millimetres, we removed data, i.e., set the respective sample to 'NA', if the sample entry was '− 1' or if it was equal or less than '0' (Step 1). For both the left and right pupil separately, we filtered the data using a percentile cut-off filter (Step 2) and linearly interpolated resulting gaps not exceeding 4 samples (Step 3). We then averaged left and right eye pupil data (keeping one value where only one was captured), then averaged and filtered the resulting array with the same procedures as above (Step 4). In a subsequent step we removed the familiarization trials samples from the dataset, excluded the 'Fixation'-element for each trial, included the type of familiarization as a separate variable (column), trimmed the dataset to remove columns irrelevant to the statistical analyses, and transposed the data to arrive at a final dataset, including all subjects, with the following columns: 'Subject', 'Session', TrialId', 'CIBLE' (The object reached for in the test trial.), 'TRAJECTOIRE' (The path reached along in the test trial.), 'VIDEO', Stimulus, 'Fam_Obj' (The object of the familiarization phase.), 'Fam_Path' (The path of the familiarization phase.), plus 721 columns of pupil data reflecting the 12 s of the movie-segment in each test trial (Step 5). In this final dataset each subject was included with one row per test trial (if test trial data were recorded).

6.1.2. Post-processing

Based on the information provided in the columns regarding the familiarization and test trials, we calculated two new variables (columns) to reflect whether a respective test trial represented a case of reaching for the expected or unexpected object (Cond_Obj;

Expected Object vs. Unexpected Object) and whether it represented a case of reaching along the expected or unexpected path (Cond_Path; Expected Path vs. Unexpected Path; Step 6). We then baseline-corrected the data per subject and trial. The baseline was calculated as the average pupil size of the 500 ms preceding the moment the hand rested on the object (2.5–3 s). From all subsequent values (3–12 s), we subtracted the baseline and divided by the baseline to arrive at time course data per subject and trial (Step 7). In a final step, these baseline-corrected pupil dilation data were averaged for the first 6 s and we additionally calculated the percentage of found samples per trial (Step 8). We included in the visualizations and statistical analyses only those trials for which at least 50% of the data were captured, i.e., had values that were not *NA*. The initial distribution of data points per condition was as follows: $n = 27$ (Expected Path/Expected Object), $n = 27$ (UNexpected Path/Expected Object), $n = 26$ (Expected Path/UNexpected Object), $n = 30$ (UNexpected Path/UNexpected Object). Following the pre- and postprocessing, the final distribution of valid data points was: $n = 11$ (Expected Path/Expected Object), $n = 11$ (UNexpected Path/Expected Object), $n = 12$ (Expected Path/UNexpected Object), $n = 11$ (UNexpected Path/UNexpected Object).

6.1.3. Statistical analysis

The main model had the following structure: ChangeInPupilDilation ~ ConditionObject * ConditionPath + TrialNumber + (1 + TrialNumber || Subject). Therefore, in addition to three fixed effects, we included random effects with an intercept for Subject and a slope of TrialNumber (z-transformed) on Subject. We calculated the *p*-value for the interaction term by dropping the interaction term and comparing the main model to a reduced model comprising only main effects, using the function *drop1()*. To calculate the *p*-values for each main effect, we compared the full model to reduced models in which the respective fixed effect was missing, using the function *anova()*. Finally, to test whether the two fixed effects of interested had a combined influence on the dependent measure, we compared the full model to a reduced model without ConditionObject and ConditionPath. In summary, we fitted a single main model and to calculate the *p*-values for the individual or combined effects, through likelihood-ratio tests, we fitted 4 reduced models. These analyses were based on 20 subjects. Visual inspection of (1) the histogram of the main model residuals and of (2) the fitted values against model residuals did not reveal an issue of violated assumptions.

6.2. Results

The change in children's pupil dilation did not systematically vary as a function of both the type of object reached for and the path reached along, LRT(interaction term): $\chi^2(df = 1) = 1.18$, $p = .28$. In addition, the results revealed no main effect of the type of object reached for, $\chi^2(df = 2) = 2.79$, $p = .25$, and no main effect of the path reached along, $\chi^2(df = 2) = 5.8$, $p = .055$ (see Fig. 14A). Finally, there was no combined effect for the type of object reached for and the path reached along, LRT(omnibus test): $\chi^2(df = 3) = 7.14$, $p = .068$. Descriptively, children showed the least change in pupil dilation in the Expected Path/Expected Object-condition ($M = 0.2\%$, $SD = 4.95\%$) and the largest increase in pupil dilation in the UNexpected Path/UNexpected Object-condition ($M = 3.77\%$, $SD = 4.16\%$). Pupil dilation change in the UNexpected Path/Expected Object-condition was similarly high ($M = 3.48\%$, $SD = 4.64\%$) and the change in the Expected Path/Unexpected Object-condition was lower yet ($M = 2.15\%$, $SD = 2.90\%$). Finally, we did not find an effect of trial number on the change in children's pupil dilation, $\chi^2(df = 1) = 1.54$, $p = .21$. Visual inspections of the time course data further suggest that while descriptive condition differences did emerge from the moment the reaching movement was completed, i.e., '0', the data remained variable, i.e., as indicated by the standard errors, over the course of the test trial (see Fig. 14B and Fig. 15).

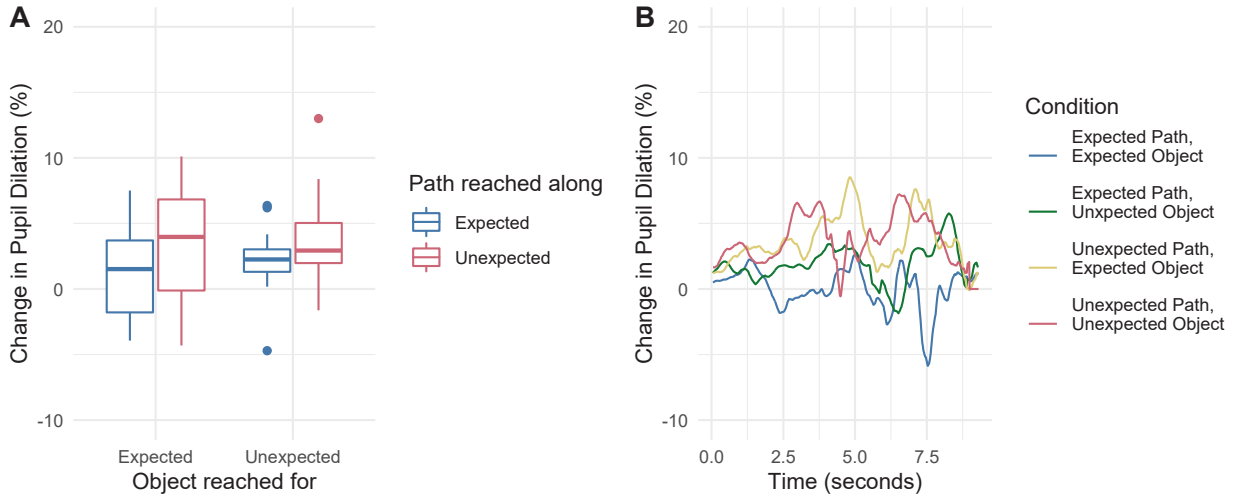

Fig. 14. Summary of pupil dilation changes averaged over the first 6 s following the completion of the reaching action (**A**). For the purpose of visualization, the time course data were additionally filtered using a moving average filter and the 6 s time window of analysis is marked by a grey 'corridor' (**B**).

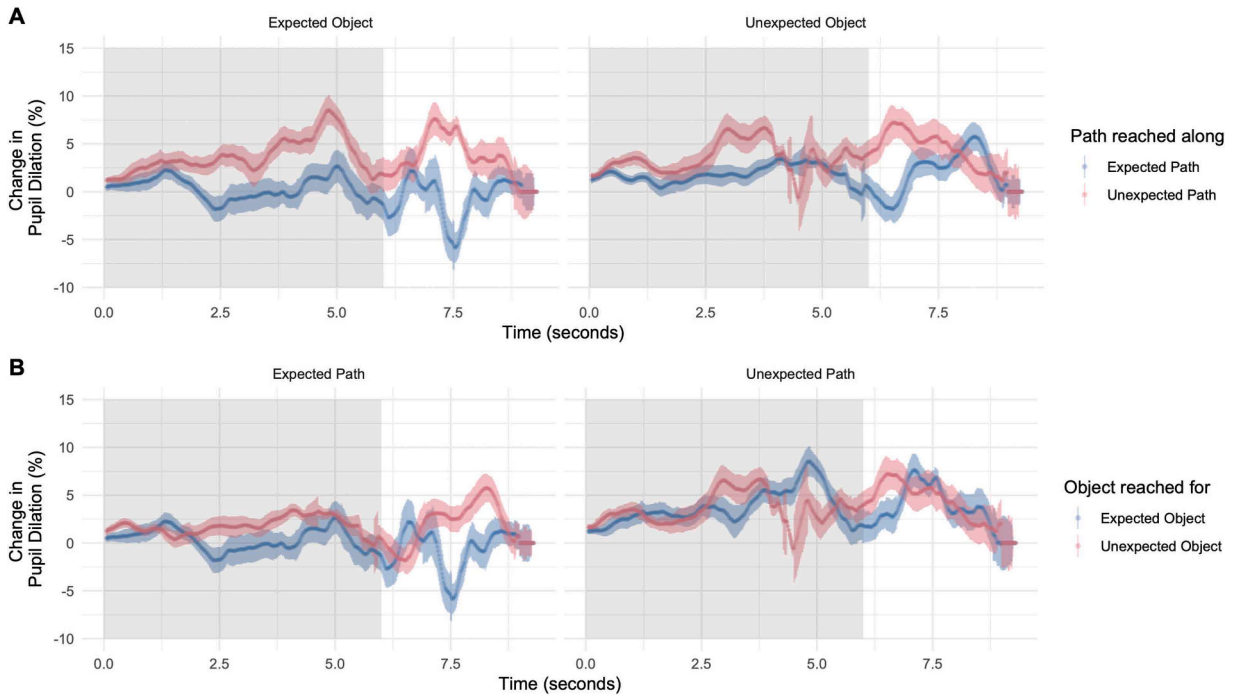

Fig. 15. The time course data of pupil dilation change averaged across subjects. The dots represent the average pupil dilation change at a given point in time while the bars represent the standard error at each time point. The grey 'corridor' marks the 6 s analysis time window across which the data were averaged for statistical analyses.

6.3. Discussion

The results do not support a strong psychological interpretation of changes in children's pupil dilation. If infants encoded the actor's, i.e., the hand's, goal then a case in which the hand reached along a novel (unexpected) path for a novel (unexpected) object would represent a goal-incongruent scenario. Descriptively, pupil dilation did increase most in this target condition following the completion of the reaching action. In addition, and based on visual inspection of the time course of pupil dilation changes, children showed stronger novelty responses if the hand reached along a novel path whereas the hand reaching for a novel object yielded a blunted novelty response. One may speculate whether infants' goal attribution encompasses two levels, one encoding the path (the motion) and one encoding the object (the target), and that pupil dilation changes reflect these different levels of encoding. However, on the basis of the current dataset none of statistical analyses conducted provided sufficient evidence against the null-hypothesis and as a consequence, the observed pattern in pupil dilation changes may not generalize beyond the current sample.

7. Results and Discussion (Cluster Mass Permutation - Team D: Hochmann)

7.1. Pupillometry Method

For this analysis, we considered only the first 5 s of each trial, corresponding to the movement of the hand towards an object and two seconds after the hand reached one of the two objects. Additional analyses on the rest of the trials did not provide interesting results due to the sparsity of the data.

We defined an area of interest (720 pi x 540 pi) corresponding to the surface of the stimuli on the screen. The pupil diameter for the left eye was recorded for fixations in that area of interest. Similar results were obtained analyzing the right pupil, or the average of both pupil diameters. In order to include a maximum of trials, and due to the sparsity of the data, we considered a long baseline time window lasting 1500 ms. The average pupil diameter in the baseline window was subtracted from all data points. Two analyses will be reported below. In the first analysis, we considered a baseline beginning 750 ms before the beginning of the movie playback and ending 750 ms after the movie playback. In the second analysis, we considered a baseline of 1500 ms before the moment the hand reached the object (3000 ms after the beginning of the movie playback.

We excluded trials with less than 100 ms (~ one fixation) of pupil diameter information in the baseline time window. In the rest of the trials, missing data were linearly interpolated. Participants lacking one trial in at least one condition were excluded from further analyses. Note that contrary to previous work (e.g., Hochmann & Papeo, 2014; Hochmann, 2022; Hochmann & Toro, 2021), we did not impose a criterion on the proportion of good pupil data over the trial duration. Applying such criterion (even a very inclusive one) would have resulted in the rejection of more than two third of the participants, yielding a sample size that would be insufficient to

conduct conclusive analyses.

Cluster mass permutation tests (Hochmann & Papeo, 2014; Hochmann, 2022; Hochmann & Toro, 2021; Maris & Oostenveld, 2007) were implemented to probe the variation of pupil dilation in response to two factors Object change (Yes, No), Trajectory Change (Yes, No) and their interaction.

7.2. Results and Discussion

First, we analyzed the data considering a baseline starting 750 ms before the beginning of the movie playback and ending 750 ms after the beginning of the movie playback (Fig. 16, left panel). Twelve (out of 30) participants were rejected from this analysis due to insufficient pupil information in the baseline, yielding a group of 18 participants.

The analysis of the time-course of pupil dilation with cluster mass permutation tests found no significant main effect of Object Change or Trajectory Change, but an interaction of the two factors in the 767–1667 ms time window; $P = .04$. To interpret this interaction, we averaged the pupil dilation in the 767–1667 ms time window. The interaction reflected larger pupil dilation when there was an object change or a trajectory change compared to no change or a change of both the object and the trajectory. In details, the object change yielded larger pupil dilation than the no-change condition ($t(17) = 2.26$; $P = .019$, one-tail); the trajectory change yielded larger pupil dilation than the no-change condition ($t(17) = 1.88$; $P = .039$, one-tail). Other comparisons were not significant ($Ps > .05$ one-tail).

Another way of interpreting these results is that pupil dilation was larger in the conditions where the objects swapped positions, as in the original Woodward paradigm, compared to the two conditions where the objects stayed in the same position as in the habituation phase ($t(17) = 3.03$; $P = .007$).

Cluster-mass permutation tests further compared the time-course of pupil dilation of respectively the object change, the target change and both-changes conditions versus the no-change condition. These analyses identified larger pupil dilation for the object-change condition compared to the no-change condition in the 1283–2217 ms times window ($P = .04$, one-tail), and larger pupil dilation for trajectory-change condition compared to the no-change condition in the 1250–1933 time window ($P = .04$, one-tail). The both-changes condition did not differ significantly from the no-change condition. Finally, a cluster-mass permutation test compared the time-course of pupil dilation in the object- and trajectory-change conditions, and found no significant differences.

Second, we analyzed the data considering a baseline starting 1500 ms before the moment the hand reached the object (3000 ms after the beginning of the movie playback) and ending when the hand reached the object (Fig. 16, right panel). Nineteen (out of 30) participants were rejected from this analysis due to insufficient pupil information in the baseline, yielding a group of 11 participants. The analysis of the time-course of pupil dilation with cluster mass permutation tests found no significant main effect of Object Change or Trajectory Change, and no interaction of the two factors. Pairwise cluster-mass permutation tests further compared the time-course of pupil dilation in the difference conditions and found no significant differences.

In sum, the results obtained here do not allow conclusions about the issues raised by the Woodward paradigm with respect to the relative weights of trajectory and goal (object) in the representation of a grasping action. Rather, our analyses suggest that the change in the position of objects triggered pupil dilation, suggesting that infants encoded the position of the objects during the habituation phase.

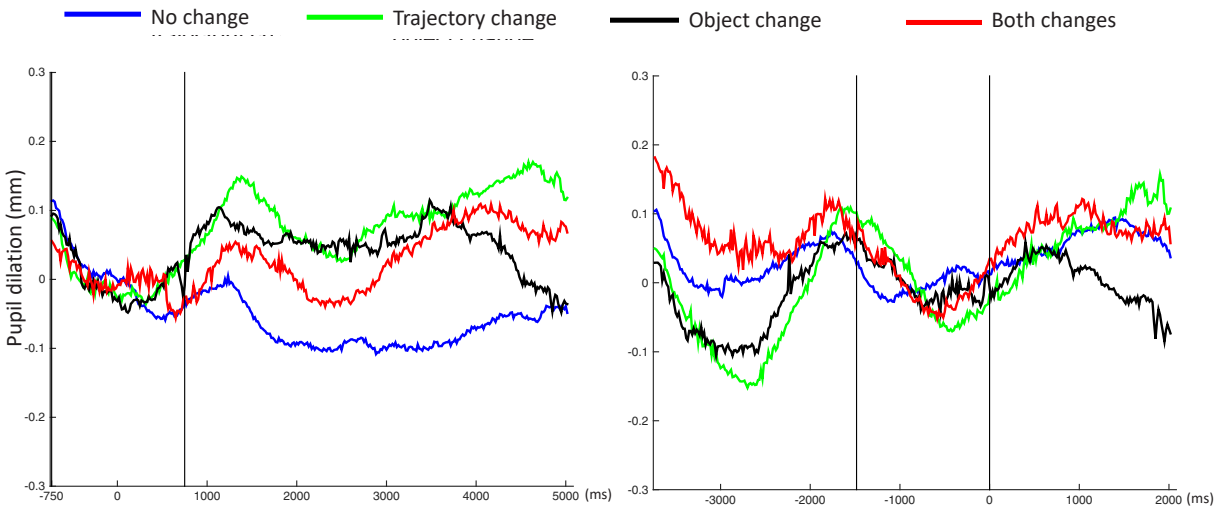

Fig. 16. Average pupil dilation for the no-change (blue), Trajectory change (green), object-change (black) and both-changes (red) conditions. In the left panel, the baseline is taken in the − 750–750 ms time window, 0 corresponding to the beginning of the movie playback. In the right panel, the baseline is taken in the − 1500–0 time window, 0 corresponding to the hand grasping the object.

One potential problem in applying our analyses to the current pupil data is that participants received only one trial per condition. This may result in high noise in the data. Our methods to analyze pupil data may be better suited for paradigms involving multiple trials per conditions.

8. Results and discussion (Linear Mixed Model - Team E: Mayer and Liszkowski)

8.1. Pupillometry method

8.1.1. Preprocessing

All data were preprocessed and analyzed using R (version 4.2.1, available at http://cran.r-project.org) using a script that was specifically written for this data set. To clean and preprocess the data, we compiled all data files into a single data frame. Preprocessing procedures included 1) data cleaning 2) creation and specification of relevant variables (e.g. condition and timestamp) 3) baseline normalization.

8.1.2. Data cleaning

Pupil measures for both eyes were compiled as follows: if both eyes were indicated as valid (validity = 0), we took the mean pupil, if not, we selected the pupil value with the lower validity rating. Pupil entries with a validity rating of 4 were later removed and excluded from analyses. Further, we excluded pupil entries if the corresponding gaze values lay outside the stimulus display area of the screen - a value we computed based on screen resolutions as described in the section 'Apparatus' (+200px cushion). Following baseline-correction (see below) we removed extreme relative pupil changes (>.5), as cognitive influences on the pupil tend to result in smaller changes (Mathôt, 2018).

8.1.3. Variables

To prepare the raw data for our analysis approach, we computed the following variables:

goal: A factor with the levels *old goal* and *new goal*. This factor is based on familiarization and denotes whether the object the hand reaches for in this trial (elephant or ball) is congruent with the object the infant was habituated to.
path: A factor with the levels *old path* and *new path*. This factor is based on familiarization and denotes whether the trajectory of the hand reach in this trial (left or right) is congruent with the trajectory the infant was habituated to.
position: A factor with the levels *old position* and *new position*. This factor is based on familiarization and denotes whether the object positions in this trial are congruent with the object positions the infant was habituated to.
timestamp: Denotes the time interval from the beginning until the end of a trial. The timestamp was computed based on sampling rate (60 Hz) and onset of each stimulus event.
outcome window: Following close visual inspection of the stimulus materials, we set the outcome window (total = 2750 ms) from 1250 ms (when the hand starts moving towards object) until 3000 ms (when the hand stops at object) plus 1000 ms still frame (of the hand on the object). To account for a time lag in the pupil dilation response, we adjusted all time windows + 500 ms (see Pätzold & Liszkowski, 2020).

8.1.4. Baseline correction

We performed subtractive baseline-correction (Mathôt et al., 2018; van Rij et al., 2019) by subject and by trial to account for luminance effects and other variability confounds. Baseline was set as the average pupil size in case of valid gaze entries during 500–2500 ms post onset (2000 ms duration) of the infant-controlled still frame fixation period prior to the onset of each trial.

8.1.5. Interpolation and filtering

In line with recommendations on pupil data in mixed effects frameworks (see van Rij et al., 2019) we opted not to interpolate and/or filter the data during preprocessing.

8.2. Results

Before running the analyses, data were segmented and processed separately for the predefined outcome window. To test the effect of goal and path on (relative) pupil outcomes, we fitted a linear mixed effects model using the lme4 package (Bates et al., 2015). Our model included goal and path (and their interaction) as fixed effects and random intercepts for each subject. Significance was calculated using the lmerTest package (Kuznetsova et al., 2017), which applies Satterthwaite's method to estimate degrees of freedom and generate p-values for mixed models. The model specification was as follows: (relative) pupil ~ goal + path + goal*path + (1| subject). Factor levels for goal and path in this model are sorted from old to new.

Model results (Table 3) show that infants' pupil sizes were significantly smaller for new goals compared to old goals ($\beta = -.06$, SE=.006, $t = -8.98$, $p < .001$) and larger for new paths compared to old paths($\beta = .06$, SE=.007, $t = -8.08$, $p < .001$). The interaction was not significant ($\beta = .01$, SE=.01, $t = .81$, $p = .419$), therefore, pupillary responses to path did not differ as a function of goal and vice versa. The model's total explanatory power is substantial (conditional R2 =.51) and fixed effect alone account for a large amount of explained variance (marginal R2 =.03). Results are visualized in Fig. 17.

Following our planned analyses, we chose to fit a second mixed effects model to the data. This approach is motivated by the current

Table 3
Regression table: Predicted pupil dilation response by goal and path.

Effect	Estimate	95% CI		*p*
		LL	*UL*	
Fixed Effects				
intercept:	-0.00	-0.07	0.06	.952
new goal	-0.06	-0.07	-0.05	**< .001**
new path	0.06	0.05	0.07	**< .001**
new goal*new path	0.01	-0.01	0.03	.419
Random effects				
σ2	0.03			
τ00 subject	0.03			
ICC	0.49			
N subject	24			
Observations	5859			
Marginal R^2 / Conditional R^2	0.032/0.510			

Note. N = 24. CI = confidence interval; LL = lower limit; UL = upper limit.

design structure: in the classic Woodward paradigm, at test the locations of the objects are swapped, that is, the object display is perceptually different from the habituated object display, which may further affect looking time and pupil sizes. The original Woodward-effect demonstrates a difference between same versus different reaching paths in the swapped object display, but it is not clear whether this would hold true for the familiar object display or interact with the familiarity of the object display. As can be deduced from Fig. 17, pupil response by path only increases in trials where objects were displayed at the new location. Thus, we included object position on the display as a predictor instead of goal: (relative) pupil ~ position + path + position*path + (1|subject). In line with predictions, the regression (see Table 4) revealed a significant interaction term between the two factors (β = .11, SE=.009,

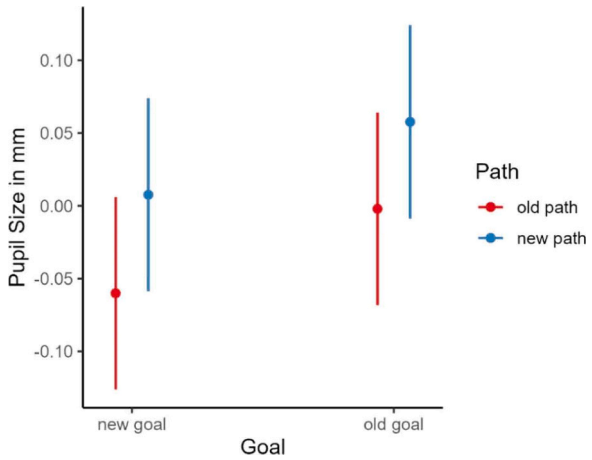

Fig. 17. Predicted relative pupil size by condition: goal and path.

Table 4
Regression Table: Predicted Pupil Dilation Response by Position and Path.

Effect	Estimate	95% CI		*p*
		LL	*UL*	
Fixed Effects				
Intercept	-0.00	-0.07	0.06	.952
new path	0.01	-0.00	0.02	.181
new position	-0.06	-0.07	-0.05	**< .001**
new path*new position	0.11	0.09	0.13	**< .001**
Random effects				
σ2	0.03			
τ00 subject	0.03			
ICC	0.49			
N subject	24			
Observations	5859			
Marginal R^2 / Conditional R^2	0.032 / 0.510			

Note. CI = confidence interval; LL = lower limit; UL = upper limit.

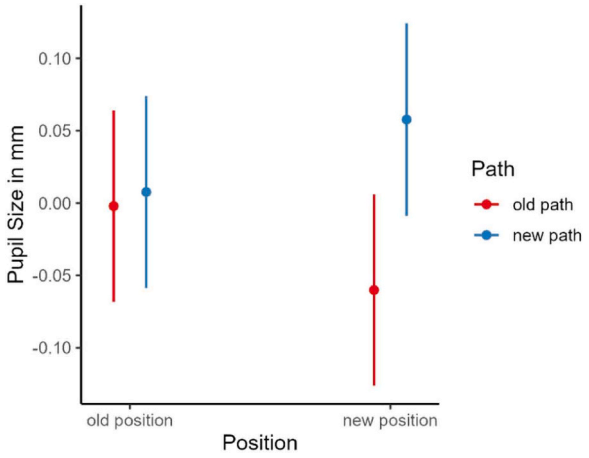

Fig. 18. Predicted relative pupil size by condition: object positions and path.

$t = 11.21$, $p < .001$).

These results reveal that only when object locations were swapped, infants' pupils dilated in response to the hand moving through the new path to the habituated object compared to the hand moving through the habituated path to a new object. However, when objects remained at their familiarized locations, there was no difference between new and old reaching paths (for illustration, see Fig. 18).

8.3. Discussion

The results we obtained from this pupillometric method could not support the findings by Woodward (1998) and others that have employed this paradigm with looking time measures (Biro & Leslie, 2007; Kiraly, Jovanovic, Prinz, Aschersleben, & Gergely, 2003). If infants' longer looks in trials where the hand grabs a new toy on the old path indicate surprise about the observed changes in goals, infants' pupil response should also reflect surprise in the current study and dilate accordingly. However, our results revealed that infants' pupils constrict in response to observing goal changes on the familiar path and dilate in response to observing path changes towards a familiar object. At a first glance, this suggests that infants attend to path changes rather than goal-orientation when observing novel events. These findings should be interpreted with caution though, as there are confounding factors worth noting. Beyond changes in path (left versus right) and goal (elephant versus ball), infants also inadvertently witnessed changes in object location on the video display. First, surprise about an unexpected object location change might lead to increases in pupil dilation at the beginning of a trial, leading to potential confounds in early pupil traces. Second, these perceptual differences may impact how infants process changes in paths. We addressed the latter by reporting an additional model that takes into account familiarity of the object display. The results clarify that throughout the experiment, infants' orientation to path only surface in their pupil response when objects are displayed at a different location than during familiarization. Consequently, we cannot firmly conclude that infants show less surprise towards unexpected goal changes or that infants show more surprise to unexpected path changes. Instead, it appears as though several factors may be affecting pupil dilation differently and it would be premature to propose that infants at this age do not understand goal attribution. It is possible that the current stimuli not only trigger action-goal processing but also object processing. Yoon, Johnson & Csibra (2008) found that 9-month-old infants encode object location at the expense of object identity when observing an object-directed reaching event. In light of this, it is plausible that infants' pupil responses in the current study reflected heightened attention to the reach when it revealed a change in the habituated location of the object.

While the parameters of the current design may be well suited for another method, it presents limitations to pupil analyses with a predefined time window. Firstly, the current video stimuli display inconsistencies in the temporal onset of action. For instance, in the video depicting a hand reach to the ball on the left, the action was completed sooner than in other trials. Further, across trials, the agent reaching for objects used her right hand to reach, therefore extending her arm differently depending on the path she took. For objects positioned on the right-hand side, the hand had to move around to the outer corner of the object before grasping the object. This raises the concern that reaching events to one side followed a more salient motion path than to the other.

Secondly, the study design did not include a neutral baseline period with a minimally stimulating display. As the first frame of the video is set as still frame/fixation period, the display shows both objects with the hand at the bottom center. In case of location changes between test trials, the baseline period might capture pupillary reactions to observed location changes and trigger encoding of spatial information. In our analyses, predicted coefficients are interpreted as pupil changes relative to each participant's baseline per trial. Therefore design-related baseline confounds may skew pupil estimates during outcome.

Thirdly, the full-factorial design duplicates within-subject conditions, thus departing from typical Violation-of-Expectation paradigms. Displaying four different test videos bears the danger of changing expectations during the test period. Further, it precludes the possibility of a sufficiently powered, balanced clean first trial analysis. Finally, the length of the test time frame, while necessary for

looking time studies, defies the great advantage of pupillometric studies to go beyond single-trial designs and test across multiple trials.

9. Results and discussion (Normalized Time Bins / Multilevel Modeling - Team F: Ross-Sheehy)

9.1. Pupillometry method

9.1.1. General approach

Pupil diameter (mm) was sampled binocularly at 60 Hz throughout each 12 s test interval, resulting in approximately 720 samples per trial. These data were preprocessed to remove invalid data, small subsequent gaps, and artifacts (e.g., blinks, looks away, noise). Data were then averaged into 500 ms bins, and baseline corrected to remove random pupil variation due to differences in overall size, arousal, and luminance adaptation (see *Data preprocessing* for details). Linear mixed effect models (LME) will be used to assess pupil change from baseline for each condition, and planned comparisons will be conducted to assess dishabituation.

9.1.2. Data preprocessing

Data preprocessing approach was adapted from procedures described by Mathtôt and others (Mathôt et al., 2018; Mathôt & Vilotijević, 2023). First, raw sample data were parsed into individual trials, and invalid data were removed (i.e., Tobii validity codes < 3, and/or looking off screen). Remaining pupil was then averaged across left and right eyes discarding any samples that contained only a single eye. Next, small gaps due to noise and/or removal of invalid data were filled with nearest values (max gap 3 samples. ~50 ms). After this, data were examined for blinks and other large artifacts. Due to the relatively large changes in pupil diameter over the course of the 12 s trial, thresholds based on deviation from overall mean diameter were ineffective. Thus, samples were examined for outliers using a moving window approach (window = 24 samples, ~400 ms). This approach was very sensitive to artifacts despite the large, slow pupil changes observed here. Samples were marked as outliers if they exceeded 3 scaled mean absolute deviations from the moving window median (i.e., *Hampel filter*). These outliers were then removed and filled using cubic spline interpolation, an approach which reduces artifacts around noisy edges, and preserves signal shape. Visual inspection of individual filtered trials revealed reasonable results (see details in MATLAB preprocessing script, and representative results available at the OSF repository of this project).

The final step in data preprocessing was to perform a baseline correction. A subtractive baseline correction was selected as it maximizes power to detect pupil changes while guarding against artifacts sometimes produced by divisive corrections (Mathôt et al., 2018). Cleaned sample data were parsed into *fixation* (fixation interval just prior to test onset to allow luminance adaptation) and *test* (12 s movie consisting of 3 s of *hand moving* toward one of the targets, and 9 s of *hand stationary* touching the target). Pupil samples during test were averaged into 500 ms bins (30 samples per bin, 24 bins per test), and the last 500 ms of fixation was averaged to create a baseline. Baseline was then subtracted from each test bin, resulting in an array of baseline corrected pupil values. For plotting and modeling, a value of 0 was appended to the beginning of the array, resulting in 25 total bins/trial.

9.1.3. Data inclusion criteria

In addition to the data validity requirements noted in *Data preprocessing* above, subjects were excluded from analysis if they did not view all 4 test trials (n = 4), or if they did not produce valid gaze for at least 1 test trial (n = 2). Periodically, individual trials had to be dropped due to the lack of valid gaze during the baseline interval. Out of 377 trials, 39 had to be dropped due to missing baselines, 13 of these were test trials (3.4%). Despite the loss of data, baseline correcting is critical to ensure pupil changes are being driven by cognitive processes elicited by the condition manipulation rather than luminance changes, arousal, or overall fatigue (Ross-Sheehy & Eschman, 2019).

9.2. Results

Pupil change from baseline was analyzed using R Studio (RStudio Team, 2015) with packages *lme4* and *lmerTest* (Bates et al., 2015; Kuznetsova et al., 2017). This approach is ideal for pupillometry, as LME models are robust to sparce and missing data, and can handle the interdependence of pupil diameter over time and across conditions (Singmann & Kellen, 2019). Target (familiar or novel) and Path (familiar or novel) were *deviation* coded so fixed effects could be interpreted as main effects and interactions.

Model fitting proceeded as follows. First, a baseline model (m^1) was created including only a fixed effect of time, and random effects of time (slope) in subject (intercept). Inclusion of these random effects allowed pupil change to vary by subject. The next model (m^2) added fixed effects of Path and Target, which significantly improved model fits, $X^2(2, N = 1360) = 10.30$, $p = .006$. The final model (m^3) added interactions for all fixed effects, which once again significantly improved model fits, $X^2(4, N = 1360) = 27.875$, $p < .001$ (see Table 5 for model estimates, Fig. 19A for mean pupil change from baseline, and Fig. 19B for plotted model fits).

Estimates were examined for the best-fitting model (m3) and revealed a marginal main effect of time, $\beta = -0.011$, $SE = 0.005$, $t(25.815) = 1.999$, $p = .056$, driven by slow pupil dilation over time. Results also revealed a significant Path by Time interaction, $\beta = -0.005$, $SE = 0.002$, $t(1351.670) = -2.193$, $p = .028$, reflecting significant pupil *dilation* when viewing the novel path (Fig. 19C). Interestingly, a significant Target by Time interaction revealed significant pupil dilation when viewing the *familiar* target, $\beta = 0.007$, $SE = 0.002$, $t(1344.248) = 2.778$, $p = .006$ (Fig. 19D). These effects were further qualified by a significant Path by Target by Time interaction, $\beta = 0.005$, $SE = 0.002$, $t(1341.462571) = -2.314$, $p = .020$ likely driven by substantial pupil *dilation* when infants viewed the *familiar target/novel path*. Follow-up simple effects with a Bonferroni p value adjustment revealed that when infants viewed the *novel path*, pupils dilated significantly more to the familiar target than to the novel target, $t(1352) = -4.769$, $p < .001$. In addition,

Table 5
Estimates and standard errors for each LME model. Significant effects for best-fitting model are indicated in bold. The Chi Squared test compared the m^n to m^{n-1}. Fit metrics included log likelihood (LL), the Akaike information criterion (AIC), and Bayesian information criterion (BIC).

	Model Comparisons		
	m1	m2	m^3
Constant	0.027	0.027	0.028
	(0.028)	(0.028)	(0.028)
Path		-0.032+	0.024
		(0.017)	(0.030)
Target		0.042*	-0.030
		(0.016)	(0.030)
Time	0.005*	0.005*	**0.005**+
	(0.002)	(0.002)	**(0.002)**
Path by Target			0.023
			(0.06)
Path by Time			**-0.005***
			(0.002)
Target by Time			**0.007****
			(0.002)
Path by Target by Time			**-0.011***
			(0.005)
X^2	–	$p = .006$	$p < .001$
Best Fit	no	no	**yes**
Observations	1360	1360	1360
LL	-334.82	-310.84	-292.04
AIC	677.64	631.68	606.08
BIC	698.50	657.75	663.45

Note: +p < 0.10; *p < 0.05; * *p < 0.01; * **p < 0.001

when viewing the *familiar target*, pupils dilated significantly more to a novel path than a familiar path, $t(1348) = 4.165$, $p = .001$. No other comparisons were significant.

Next, planned comparisons were conducted to determine if infants dishabituated to any of the three novel test conditions relative to the familiar test. To accomplish this, separate two sample t-tests were conducted comparing average pupil for the *familiar target/ familiar path* to each of the three test conditions (equal variance assumed). The only condition that differed significantly from the familiar test was the *familiar target/novel path* condition, producing significantly larger pupil dilations than the *familiar path/familiar target* condition, $t(357) = -2.42$, $p = .016$. Neither the *novel target/ familiar path* (p = .644), nor the *novel target/novel path* ($p = .701$) differed from the familiar test.

9.3. Discussion

The goal of the present task was to determine if when habituated to a goal-oriented motion (a hand moving along a *path* toward an *object)* infants encoded the *action* (i.e., movement path), the *target* of that action (i.e., the toy), or the *goal* of the reacher (i.e., to *retrieve a particular object*). Test trials were fully crossed manipulations of Target (ball or elephant) and Path (left or right) resulting in four distinct 12 s movies: *familiar target/ familiar path*, *novel target/familiar path*, *familiar target/novel path*, and *novel target/novel path.* Previous work using looking time is somewhat mixed, with some results favoring a "goal attribution" interpretation (Woodward, 1998, 2009), and others favoring an "encoding/information processing" interpretation (Jackson & Sirois, 2009, 2022). Though somewhat mixed, our results favor the latter. If infants were sensitive to the goal of the reacher (i.e., to *retrieve a particular toy*), then we would expect the greatest pupil dilation when the hand reached to the new target, regardless of the path. However, we found the opposite effect; pupils dilated significantly more when the hand reached to a *familiar* target. Although it could be argued that in this context, pupil dilation might reflect recognition (e.g., Ross-Sheehy & Eschman, 2019), we would expect this pattern to be consistent across conditions and it is not; pupil shows significant dilation in response to path *novelty*. In addition, if infants represented the goal of the reacher, we would expect infants to dishabituate to the novel target/familiar path condition, as in Woodward (1998). Instead, planned comparisons revealed no difference from habituation.

When assessing the encoding/information processing hypothesis, results are equally puzzling. Assuming additive or even multiplicative effects of novelty across both Path and Target, we might predict that infants show moderate pupil dilation when *either* path *or* target is novel, and strong pupil dilation when *both* path *and* target are novel. However, inspection of Fig. 19 reveals that clearly is not the case. Infants seem to respond strongly to the *novelty* of the path, and the *familiarity* of the target. It is currently unclear what is driving these distinct pupil responses, or even if the effects are underpinned by the same cognitive processes. Of course, the same could be said of looking time measures, with any number of cognitive processes contributing to gaze behaviors.

Planned comparisons conducted to examine habituation effects revealed only the *familiar target/novel path* differed from habituation. Neither the *goal attribution*, nor *encoding/information processing* hypotheses can fully account for this finding. Although one might hypothesize that infants were "surprised" to find a familiar target at the end of a novel reach, there are several more likely explanations

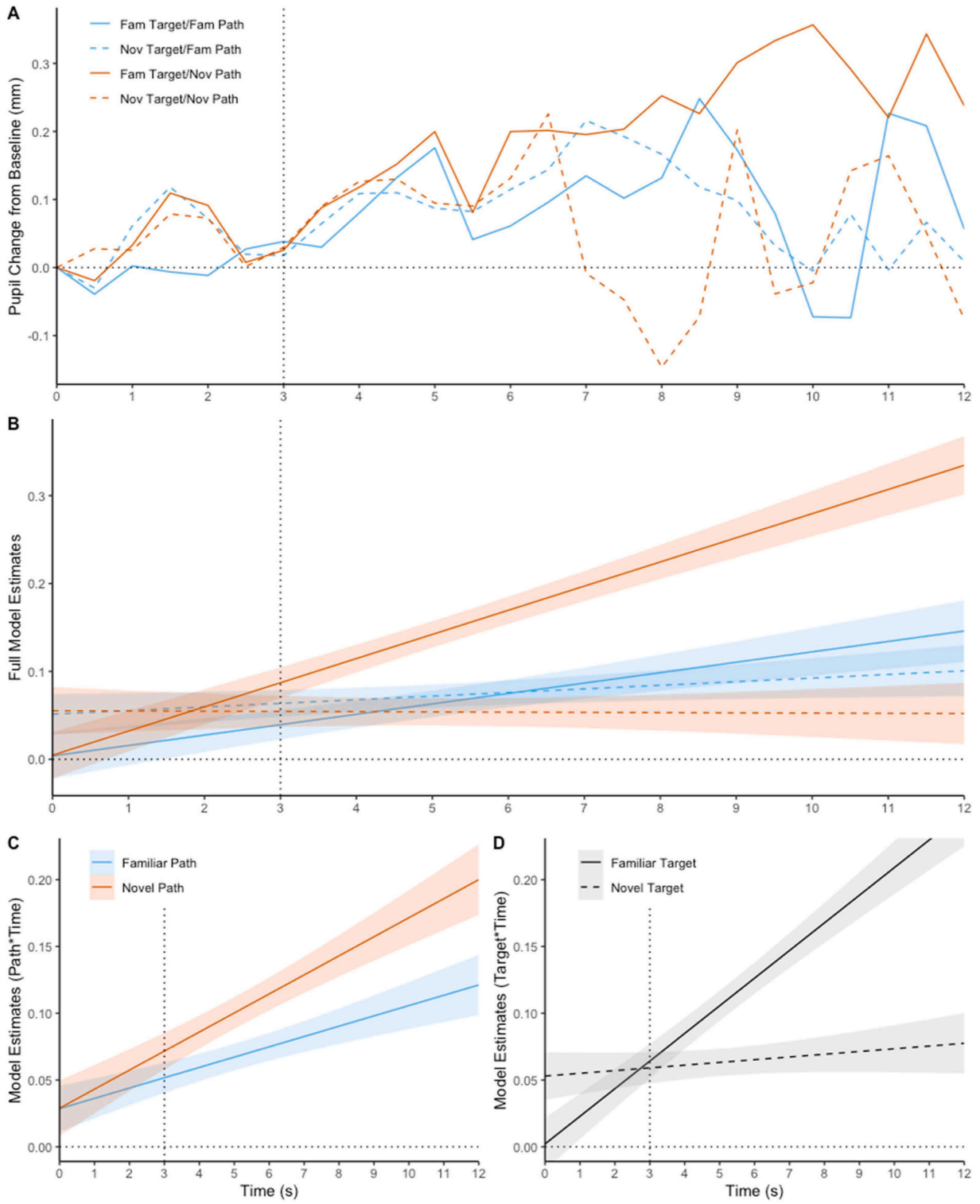

Fig. 19. Mean pupil change from baseline (Panel A), fitted model estimates for the best-fitting model (Panel B), and significant Path by Time (Panel C) and Target by Time interactions (Panel D).

for this mixed pattern of results: a) pupil does not adequately capture cognitive processes tested here, b) infants do encode action, target, or goal, but these memories subsequently wane over the course of the test trials, or c) data presented here simply lack the power necessary to differentiate our condition effects. Habituation tasks by definition are difficult, and pupils are very sensitive to eye movements and blinks. Data cleaning procedures can help, but no procedure is perfect, and even well-intentioned efforts occasionally produce artifacts. Thus, a cautious interpretation of these results is that when habituated to a hand reaching for an object, infants can differentiate a path change, as long as the object of the reach does not change.

10. Results and Discussion (Functional Data Analysis - Team G: Sirois and Brisson)

10.1. Pupillometry Method

This approach to pupillometry analysis is based on functional data analysis (FDA, Ramsey & Silverman, 1997) and has been used successfully on a wide range on infant data (Jackson & Sirois, 2009, 2022; Geangu et al., 2011; Gustafsson et al., 2015, 2016; Sirois & Jackson, 2012). A detailed introduction to the use of FDA for pupillometry can be found in Jackson and Sirois (2009), and a walk-through (with link to supporting data and Matlab code) is available in Sirois & Brisson (2014). In essence, transforming data into functions allow us to carry out analyses on the functions themselves, the results of these analyses being functions as well that can be expressed and examined over time. We can thus look at whether and also *when* significant effects are observed, without a priori arbitrary decisions about critical time windows selected for analyses.

Pupil diameter was recorded from both eyes, and we used the first 705 samples (11.75 s) from the beginning of the movie on each trial for the analyses.[2] Missing pupil samples (e.g., eye blinks, head turns), coded as − 1, require interpolation so that the trial can be used for functional analysis. Data were first filtered using a 4 Hz low pass filter, which was applied twice (forward, then back) to avoid phase drift. The typically high correlation between left and right pupil diameters allows the use of the diameter of the other eye when data from only one are missing. In cases where data were missing from both eyes, the data were linearly interpolated between the average of the last 3 valid samples before the break and the next 3 valid samples. When data were missing at the very beginning or the very end of the sample, the average of all valid samples for the trial was used as the start or end value, respectively, for interpolation. The average pupil diameter of the filtered and interpolated data from both eyes is used for further analyses.

We used the 6 samples (100 ms) immediately before playback as a baseline for each trial. The average of these 6 samples, further averaged from both eyes, was subtracted from the next 705 samples. The analyses thus examine the change in pupil diameter over the different types of test trials.

Four infants did not have data for the fourth and final test trial and were dropped from further analyses. A further 12 infants did not have a valid baseline on one or more test trials and were also excluded from analyses. We thus have 14 infants with valid data on all 4 test trials for the analyses.

To carry out functional analyses, we used B-spline functions of order 4 with 18 bases. These create smooth piecewise cubic curves (i.e. the knots that link cubic segments of the curves share slope and acceleration values). We used 18 bases because this value retained the main features of the raw data (shown in Section 3.2) yet provided substantial smoothing. As discussed and illustrated elsewhere (Jackson & Sirois, 2009), the functional analyses are robust to variations of this arbitrary parameter.

10.2. Results

The mean baseline-corrected change in preprocessed pupil diameter, averaged across infants for each problem type, is shown in Fig. 20. The same data, transformed into b-splines, are shown in Fig. 21. Comparison of the two figures highlights that while transforming data into functions introduces additional smoothing, the main features of the individual conditions, and the relative differences between conditions, are maintained.

The functional data was analyzed with a repeated-measures ANOVA with Target (familiar or novel) and Path (familiar or novel) as within-subject factors. The result of this analysis, a functional F ratio expressed over time, reveals a significant interaction between Target and Path (shown in Fig. 22). Specifically, between 1.4 and 1.77 s into the video, when the hand is following a path towards one of the toys.

We follow this significant interaction with a comparison of the effect of Target (novel - familiar) for each type of path (familiar or novel) as tests of simple effects. Fig. 23 shows the outcome of the two functional t tests used for this comparison.

As Fig. 23 shows, there is significantly more dilation when the hand follows a novel path to the familiar target than to a new target. This effect is present during a similar temporal window as the interaction, and also later in the trial for a more sustained duration. When the hand follows a familiar path, there is no difference whether the target is a familiar or novel toy.

10.3. Discussion

The results of the functional analysis of pupil diameter change revealed an interaction between Target and Path, suggesting that infants do not primarily respond as a function of goals. Furthermore, this interaction appears disordinal. There is no difference between

[2] Not all records had exactly 720 samples from the 12 s of video playback. To optimize the 4 Hz low pass filter we apply to data acquired at 60 Hz, we used the next lowest multiple of 15 samples (60 Hz ÷ 4).

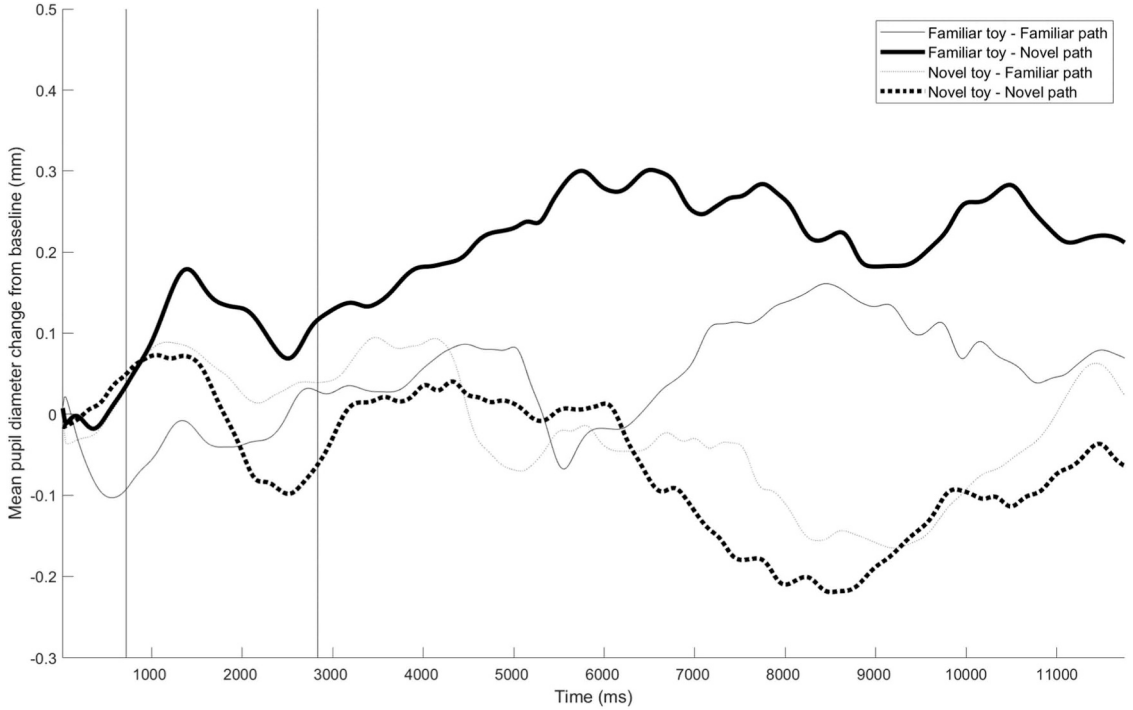

Fig. 20. Changes in pupil diameter from baseline, averaged for each test condition. The two vertical lines represent the start (left) and end (right) of hand motion, averaged between video sequences.

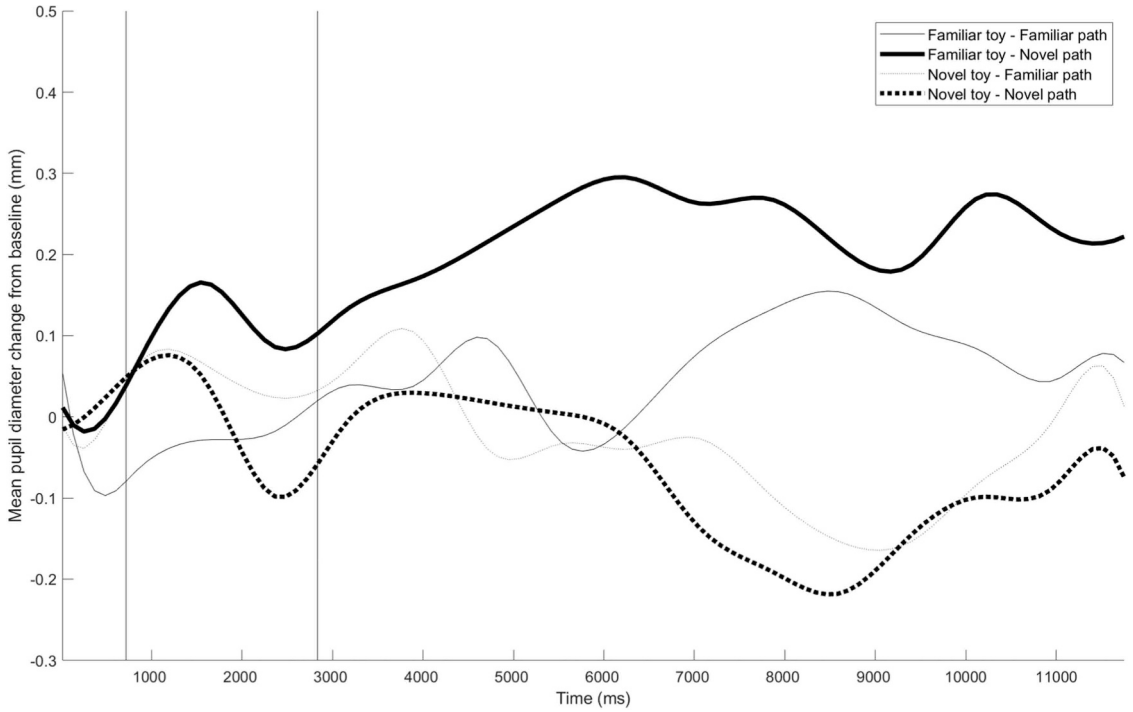

Fig. 21. Changes in pupil diameter from baseline, transformed into functional data (i.e., b-splines), averaged for each test condition. The two vertical lines represent the start (left) and end (right) of hand motion, averaged between video sequences.

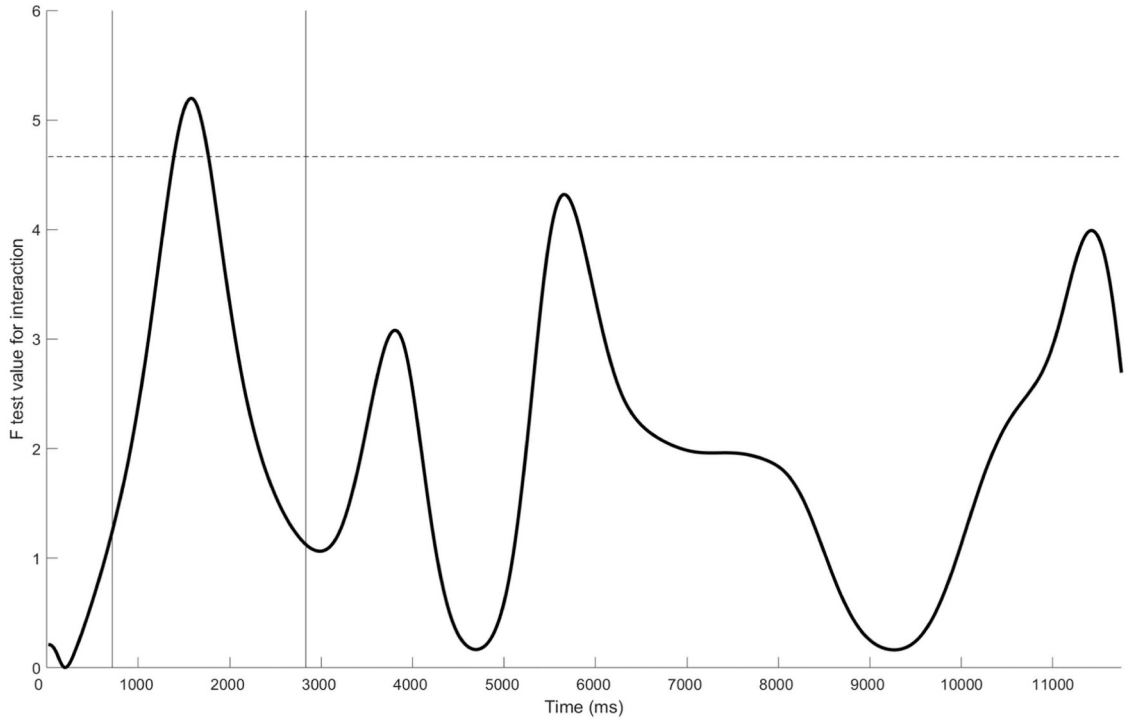

Fig. 22. Functional F test of the interaction between Target and Path on pupil diameter changes from baseline. The horizontal dashed line is the critical value for $F_{(1,13)}$. The two vertical lines represent the start (left) and end (right) of hand motion, averaged between video sequences.

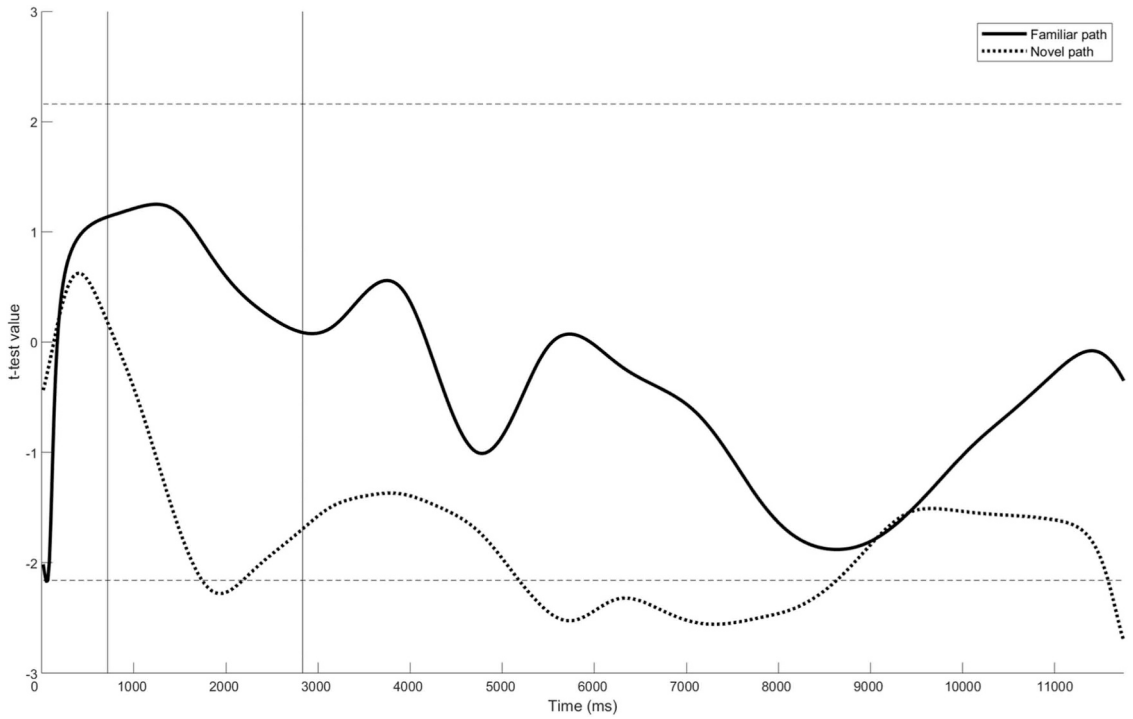

Fig. 23. Functional t tests comparing Target (novel – familiar) for each type of Path. The horizontal dashed lines are the two-tailed critical values for $t_{(1,13)}$. The two vertical lines represent the start (left) and end (right) of hand motion, averaged between video sequences.

targets when the hand follows a familiar path. However, there is only a novelty effect for Path when the hand reaches for the familiar target, during the interaction but also sustained later during the trial when the hand rests on the target toy.

The simplest conclusion from this analysis is that infants learn the association between hand and familiar target during habituation, but it is when the spatial nature of this association is disrupted at test, rather than the mere identity of the target (i.e., goal), that they react most. Indeed, the interaction is significant during hand motion, prior to the hand resting on the target toy. We cannot conclude from this analysis that infants encode goals. We can, however, stress that pupillometry and analytical tools that capitalize on the temporal unfolding of the data can reveal fine-grained aspects of infant information processing that elude more commonly used cumulative looking times.

11. General Discussion

Looking time data failed to replicate the original findings from Woodward (1998). As such, and on this metric alone, our results are inconclusive and do not support goal attribution in infants. Looking time ratio analyses are not incompatible with Woodward (1998), in that infants prefer to look at the familiar location. The factorial design of this study however highlights that this preference is irrespective of which target is at that location, whereas (and crucially) it was always the novel target toy in the original study, a confound for the goal attribution interpretation. This is further supported by the gaze data (see Fig. 6), where a familiar motion path to a familiar location yields relatively more robust looking. Incidentally, in the original study, this was systematically linked to the unexpected goal.

A key set of questions in the current paper were whether pupillometry could provide a different interpretation of cognitive processes infants use in this task, and whether such interpretation would be robust to markedly different methods. To this end, seven teams carried out their prefered analytical methods independently of each other and, crucially, blind to results from other teams. We now examine whether and how these approaches converged or diverged,

11.1. Comparison of pupil analysis approaches

Table 6 summarizes some key features of the work of the seven teams who analyzed the dataset. We notice important variations in the valid number of participants retained, the size (and position) of the pupil baseline window, and how the 12 s of data recordings per trial were used. Two teams used data from the whole trial for their analyses, one team averaged data in 25 sequential temporal bins, and four teams used specific temporal windows (each with a different position and duration, two of which did not overlap at all with respect to which portion of the trial was analysed). All but one team used some form of interpolation for missing values, the most common form being linear interpolation.

With respect to the research questions, three teams reported a main effect of path (familiar vs novel), whereas three did not (and one reported a simple effect following a significant interaction with target). One team reported a main effect of target (familiar vs novel), five did not, and one abstained because of an interaction (the difference between targets was examined as a function of path). One team reported an ordinal interaction, and three teams reported a disordinal (i.e., crossover) interaction, although of these one team found it only for one of two temporal locations from the baseline.

Overall, the most common interpretation of the data is that infants react mostly to a new path at test, especially if that new path leads to the familiar target toy. This conclusion is not unanimous, but is in line with the conclusions of Ganglmayer and colleagues

Table 6
Summary of methodological differences between teams and resulting main findings.

Team	Valid n	Baseline duration	Whole trial analysis?	Missing values	Path main effect	Target main effect	Ordinal interaction	Disordinal interaction	Compatible with goal attribution?
A	23	500 ms	1000–3000 ms	Linear interpolation	No	No	No	No	No
B	12	None, 100 ms, 200 ms	Yes	Maximum likelihood estimation	Yes	No	Yes	No	No
C	20	500 ms	3000–9000 ms	Linear interpolation	No	No	No	No	No
D	18 or 11	1500 ms*	0–5000 ms	Linear interpolation	No	No	No	Yes/No	No
E	24	2000 ms**	1250–4500 ms	No interpolation	Yes	Yes	No	No	No
F	24	500 ms	Yes, binned	Cubic spline interpolation	Yes	No	No	Yes	No
G	14	100 ms	Yes	Linear interpolation	Yes***	N/A***	No	Yes	No

* temporal location of baseline different in two analyses, leading to different participant rejection rates
** maximum duration, could be less based on available data on given trials
*** in this approach, main effects are eschewed in favor of simple effects in the presence of a significant interaction

(2019) who suggested that paths more than goals drive infant behavior. Indeed, none of the seven teams suggested that their findings were compatible with goal attribution (with some caveats). Of course, as is the case for looking times, failing to find support for goal attribution does not preclude the ability of infants to attribute them. However, unlike for looking times, pupillometry can provide a finer grained analysis of information processing in such tasks. Even when different methods are used. The three teams that report an effect of path as well as an interaction between path and target all used whole trial data and relatively shorter baseline durations. These two methodological considerations may warrant special consideration in future work (see Section 11.3).

The issue of timing differences between stimulus videos raised by Team E is another issue worthy of further scrutiny (as are the slightly different paths to reach left and right targets). Although pupil dilation is expressed relatively slowly over time, such timing or path differences can, at minimum, increase noise in the data (by misaligning pupil expression between trials), which could be more detrimental for some approaches relative to others. It is worth noting that any video sequence could be selected for habituation, and as such each sequence provided a different combination of path (new or familiar) and target (new or familiar) between infants. Consequently, this would indeed make the data more noisy, but not in a systematic way (and as such is not a confound). The fact that a majority of teams found significant effects with pupil dilation despite such noise, and in the absence of clear effects relating to looking times, certainly highlights the unique benefits of including pupillometry in infant studies.

11.2. Reflections on collaboration

In this section, the different teams were asked to share insights from taking part in this unusual collaborative endeavor. As we felt that this would be an excellent complement to other collaborative initiatives that have emerged in recent years (e.g., ManyBabies in our field), the novelty of the approach was possibly fraught with unexpected challenges. In order to make such papers more likely (if not common!) in the future, remarks provided by individual teams are reported here. As some of these remarks were shared by different teams, some teams elected not to duplicate reflections.

Team A (Blaser, Kaldy, and Donenfeld) thought that this project was an important step toward two complementary goals: 1) identifying best practices, and thereby greater standardization, in pupillometry pipelines and, 2) robustly evaluating a target hypothesis (in the present case, goal attribution). For future projects, there are five points that they think are worth considering:

1. Though this was not possible in the present project, it would be useful for teams to provide input into the study design itself before data collection. Ultimately, a team's results will be used to reject or corroborate a target hypothesis, but if teams were not involved in design, they may have concerns (statistical, like concerns about power, or paradigm-related, like concerns about stimuli or contrasts) that undermine strong conclusions.
2. To reduce variability between teams that is not relevant to the goals of the project, uncontroversial pre-processing of the data (e.g., eliminating known spurious measurements, synchronizing samples to a common timeline) should be done before sharing.
3. Teams could be provided with a common set of guidelines for the description of their pipeline. If teams can agree, to the extent possible, on technical terms and subsection organization, and are careful to explicitly specify and justify (whether on statistical or physiological grounds) substantive steps in their analyses, it would facilitate the reader making comparisons. This becomes especially helpful in determining where, and why, pipelines' results may differ.
4. Relatedly, without compromising team flexibility, teams could be provided with a common set of prescribed analyses, perhaps in three categories: a) simple calculations on which all pipelines should agree (e.g., "provide the mean of raw pupil values across both eyes for all subjects during stimulus presentation"), providing internal consistency checks, b) basic summary statistics and visualizations which would be useful in contrasting elements of pipelines that are shared, but may have differing underlying calculations and/or parameters (e.g., "graph a representative trace, with variance indicated, over all participants for each condition"; "provide a histogram of the pupil during your baseline period for each condition"), and c) final hypothesis testing (e.g., "Should we consider there to be an effect of factor X? If so, does this effect depend on the level of factor Y?").
5. Finally, the project should pre-register a method for reconciling the hypothesis-testing results of the pipelines. The goal would be to specify a 'meta-analytic' method that would allow for a formal, consolidated takeaway message.

Team B (Calignano, Russo, and Valenza) notes that in the last decade, the reproducibility crisis in psychological research has been widely recognized. Many have been the call for action resulting in promising approaches to experimental research aiming to improve replicability and transparency (e.g., Open Science Collaboration, 2015). In this framework, crowdsourced approaches aim to maximize the generalizability of results by involving independent research laboratories in experimental studies across the world (e.g., the Many Labs project series; Ebersole et al., 2016; Klein et al., 2014; the Many Babies project; Frank et al., 2017). Sharing data collection (within or across labs) inevitably leads to increasing resources, especially considering the costs and challenges of recruiting infant participants. Therefore, maximizing the datasets collected becomes crucial.

In the present collaboration, crowdsourced approaches have been extended beyond data collection to data management and statistical analysis (e.g., Silberzahn et al., 2018). In particular, the present study allows us to estimate the reproducibility of plausible statistical analysis pipelines, and the replicability of the results of a collected dataset of pupillometry measures, by asking different research teams to apply their selected statistical approaches. Therefore, a collaborative, reliable, and transparent work fashion was followed through the study. To deal with the consistent degree of freedom in the preprocessing steps prior to data analysis, our team contributed by adopting a multiverse approach (Steegen et al., 2016). Specifically, this method allows for weighting the effects of different reasonable choices, made while cleaning and organizing data (e.g., filtering, baseline correction, and areas of interest), on the resulting robustness of results (Steegen et al., 2016; Dragicevic, et al., 2019). The result of a multiverse analysis is not a single, best

method for data management, it offers transparent outcomes of several plausible choices that can affect results. Therefore, this method can significantly contribute to the existing movement for open scientific practices (Chartier et al., 2018). Yet, the different degrees of freedom selected on each dataset in the multiverse can lead to different conclusions. Here is where a crowdsourced approach can make the difference, by enriching every single-lab analysis with collective comparisons, as done in this study.

Team B also notes that some criticisms as well emerge from this practice. For instance, successfully synthesizing and discussing the results from such a rich corpus of analysis can be difficult. Specifically, the risk is for a considerable amount of work from individual teams to be in vain when it fails to serve the overall synthesis. Accordingly, orchestrating a crowdsourced project is undoubtedly challenging. Furthermore, every team must be open to discussions and ready to put the value of the collective product above the single contributions. Another concern might regard the recruitment of labs participating in the crowdsourced project. Indeed, calls for participation should travel through networks and platforms widespread around the world to really diversify the approaches and techniques recruited in the project. Lastly, working within a blinded procedure might be difficult, especially when comparing approaches against well-established practices. At the same time, crowdsourced and multimethod approaches offer the opportunity to review with unusual scrutiny the entire workflow, methodologies, especially when sharing datasets and analysis code. Therefore, even if various aspects can be further improved, crowdsourced and multimethod approaches can be useful strategies to deal with the replication crisis in psychological research. Moreover, since not all resources from labs can be directed to replicate previous findings, it might also be useful to invest in crowdsourced and multimethod approaches for statistical analysis on open datasets available online. Hopefully, this work will lead to the definition of guidelines and best practices to move the research community toward a collective, transparent, and reliable way of doing collaborative data analysis in infancy research.

Team C (Hepach) was and continues to be enthusiastic about the present approach. In its present form, the contributing teams did not so much collaborate but rather worked in parallel to pursue the same goal(s). One consequence of each team being unaware of the other teams' analyses is that the results, both in presentation and in conclusion, appear more disjointed than may necessarily be the case. Greater cohesion in the presentation of results could possibly be achieved if teams in a subsequent step, or as part of the initial commitment, agreed on labeling, coloring, plot types, and how to visualize variability (through confidence intervals, standard errors, etc.). In addition, based on the relatively small sample size some of the analytical approaches are likely statistically underpowered (as noted also by Teams A, E, and F) to detect medium to small effect sizes. It is therefore possible to arrive at a scenario where a particular hypothesis test did not yield $p < .05$ for all teams but where in fact descriptively each team's results point in the same direction. One way to address this could be to ask each team to answer specific questions in addition to testing hypotheses: *Do infants expect agents to have the goal to reach along the familiar path? Do infants expect agents to reach for familiar objects?* The analyses of each team may yield, descriptively, similar answers even if, statistically, the pattern does not fully justify rejecting the respective null hypothesis. Finally, and to echo other reflections in this section of the paper, studying the time course of pupil dilation changes can offer insights into the time course of how perceptions, or representations, of others' goals are formed. The goal attribution of 'reaching along a familiar path' versus attributing the goal of 'reaching for a familiar object' may be reflected in different temporal profiles of pupil dilation changes. Harnessing this potential of pupillometry will require teams, beyond the present paper, to orchestrate efforts toward addressing fundamental methodological questions such as *when* (after stimulus onset) and *for how long* (in terms of analysis window length) to capture psychologically induced changes in pupil dilation (see Hepach, forthcoming).

Team D (Hochmann) fully supports the implementation of this type of collaboration on a more regular basis. They would however prefer to be involved in the design of the experimental paradigm, as designs and analyses cannot and should not be conceived independently. With respect to the utility of pupillometry, and its complementarity with looking time measures, it may be interesting to work in the future on a dataset that provides unambiguous and reliable results from looking times, asking whether pupillometry provides congruent data. A caveat, however, is that looking time studies and pupillometry studies may have different constraints, leading to different designs being better suited to each of the methods. Overall, this collaboration was a very good experience.

For Team E (Mayer and Liszkowski), methods including sample size, study design, stimuli, and presentation typically predicate analytic choices. Joining a group of collaborators before learning about the methods made it difficult to register a suitable analytic approach. Thus, it may be more advisable to include multi-lab parties at the design stages of a study already, because analytic methods should not be treated independently of the study design. For example, the choice of baseline was difficult because the still frame fixation was confounded with differences in the perceptual display, and a subsequent time window was confounded with differences in visual-spatial displays. Combining the current multi-analytic approach with a multi-lab approach (e.g. as in ManyBabies) may be a promising future endeavor.

Team F (Ross-Sheehy) applauded the group approach both as a means to examine the utility and robustness of pupillometry data, and as a next-step toward refining a set of recommendations and best practices for infant pupillometry research. An additional strength of the paper was its reliance on an "naïve observer" approach; that is, analysts who were proficient in pupillometry data collection techniques but agnostic with respect to the individual research hypotheses. Although results were fairly consistent across teams, it is notable that many decisions influenced both the *observation* of effects, and the *interpretation* of those effects. These included not only general analytic approach, but also preprocessing decisions like inclusion criteria, the use of baseline correction, decisions about data validity (looks away, "invalid" looks, looks with only a single eye, etc.), artifact definition and rejection criteria, decisions about interpolation, smoothing, window size, and so on. Although the "naïve observer" approach used here helped guard against subtle biases that could impact these analysis decisions, it also made data interpretation slightly more difficult. For example, Team F noted that being unable to view the subject videos added a degree of uncertainty regarding data quality, and likely contributed to their adoption of relatively strict data inclusion criteria (e.g., pupil for both eyes, valid looks only, artifact rejection, etc.). Of course in a perfect world, data would be robust and effects would be obtained across all permutations and processing decisions. Sadly, perfect infant eye tracking data have yet to be obtained!.

Team G (Sirois and Brisson), as originators of this project and providers of the dataset, were in a different position relative to other teams. They had an a priori interest in the specific research question, and therefore could arguably be more biased than more naïve collaborators on that specific issue. Moreover, some of their previous work informed methodological decisions that modified this study relative to Woodward's (1998) seminal work. As other teams noted in their results and discussion and reflections sections, this is different to how they may have set up the study themselves. Non-trivial is how study design must align with analysis plan, which was possibly best suited for Team G. Indeed, the initial plan for this dataset was an individual paper and not this collaboration!.

While Team G knew that the data were noisy, they also knew that their analyses provided an interesting alternative to a goal attribution interpretation, which they believed to be an important contribution to discussions about the nature of the developing social mind. However, sharing raw data for alternative analyses and interpretations is a very different experience to submitting a single team paper! Aside from the unusual clerical work, there is a period of heightened anticipation between sharing data and relevant details, and reading the outcome of the work from the other teams. Thankfully, the process yielded a useful combination of commonalities and divergences that can contribute to ongoing discussions about best practices in the analysis of pupillometry data. It is a useful measure, and better ways to understand how it can inform theory are always welcomed.

As noted in the previous section, the issues of minor timing differences between videos and path differences between left and right locations were considered sources of potential noise in the data. At the design stage, prior to this collaboration, Team G considered these unavoidable. If the same hand, initially between two toys, is to grab a toy in the same way regardless of its location, the path will look slightly different (unless the hand were to rest on top of the toy, with the arm obscuring it). Likewise, the sequences were filmed using a metronome to reduce differences as much as possible between stimuli, yet differences remained. However, the fact that these differences (path and timing) were not systematic but randomly distributed (as each sequence could be any of the four test conditions between infants) were considered tolerable. If there were genuine effects path or target (or an interaction between them), they should be revealed despite the noise created by these differences. Of course, Team G knew that their method was robust to noise, but so appear to be the majority of other methods based on this collaboration. It remains that the collaboration was an enlightening experience for Team G. As science becomes increasingly open, including the sharing of datasets, we (as discipline) need to think differently about reproducibility. Sharing our data and code for analyses increases transparancy in a useful way. Our conclusions however will be far more robust if the same data yields similar conclusions from alternative analytical methods. And thus, at the design stage itself, we should plan studies that, as before, suit our preferred method of analysis, but also make it possible within reason for alternate methods to scrutinize the data as well.

Would Team G do it again? Yes, absolutely. Regularly? No, and not because of the additional work. They would rather join efforts initiated by others regularly, and occasionally initiate them. We (as a field) should be taken out of our analytical comfort zone more often to better understand the limits of the tools we regularly use within the confines of their natural home (i.e., our typical methods). This is an important opportunity for advances.

11.3. The next steps

The present novel collaboration is the first to use different analytical tools for pupillometry data from a single dataset. The most important conclusion from this is that in the case of independently authored papers, compared to such collaborative efforts, the conclusions derived from a dataset will reflect, to a degree, the choice of analysis, possibly more than inherent features of the data. Of specific interest are the duration and temporal location of baselines, as well as the temporal segmentation of trial data and the retention rates of trials and/or participants that meet inclusion criteria of the different methods. Yet, how can we make sure that a single pupillometry dataset, typical of research carried out by independent labs, provides a robust account of the phenomenon under investigation, rather than a partial (and potentially incorrect) perspective that reflects primarily statistical decisions? While this paper reports a first foray into this question, we see two immediate avenues that could provide additionnal answers.

First, as Team C observed, one way forward would be to use the current paradigm with a larger sample of infants to replicate and further substantiate the results reported here. Multi-lab data collection efforts, such as those pioneered by the Manybabies-consortium, along with multi-lab collaborative data analyses, such as the one reported in the current manuscript, are exciting avenues for the future of infancy research. Team E further suggested that these two approaches could be combined. Indeed, the planning of large-scale collaborative projects likely involves researchers with different statistical interests and toolboxes. This creates a unique and natural opportunity to plan data collection and pre-processing in such a way as to facilitate parallel, complementary analysis streams. Such projects obviously require consensus from a large number of people on myriad details (hypotheses, stimuli, procedures, dependent variables, etc.). We suggest that the consensus for analyses should be about which set of procedures rather than which one. Of course, the aim should not be to increase the likelihood of significant findings (and there should be safeguards to that end, including pre-registration), but rather to ensure that an ambitious project is not trapped (or failed) by a restricted analysis plan (Calignano et al., 2023). This is particularly relevant as some collaborators in the current project reported that their input about data collection would have been beneficial to their ability to carry out the analyses. Such large-scale projects, regardless of the research questions, would moreover be ideally suited to explore the important methodological questions of *when* (after stimulus onset) and *for how long* (in terms of analysis window length) raised by Team C. They would also be an excellent testing ground to assess the contribution of baseline (duration and location) and trial (whole or partial) to the conclusions of the different methods, which were highlighted in the current project.

Second, and based on comments from teams in Section 11.2, the ability to compare methods and identify best practice would require a modification of the approach used in the current paper. Our aim was to use a real, typical dataset such as produced by infant labs, and assess whether and how methods may converge or diverge, moreover without a sense of competition between labs to

minimize the contribution of goals (e.g., significant findings) on analysis decisions (Silberzahn et al., 2018; see also Appendix). This approach had two limitations. On the one hand, only one team was involved in the design of the study, which made the data and sample size variably suitable for the requirements of the other teams (and unsuitable for a candidate 8th team). This situation does not create an ideal scenario for comparisons (especially explaining divergences). On the other hand, because this is real data from a small, noisy sample, we do not have benchmark parameters to comparatively assess the different approaches. It can be argued that an iterative process (e.g., Silberzahn et al., 2018) with such a dataset could create a herd effect, with no guarantee that the majority findings are correct yet creating artificial confidence. In this paper, we have no guarantee that summary conclusions are objectively correct, which makes it challenging to use them as a reference to identify best practices. We thus suggest that a follow-up project should involve teams at the initial phase that determines the nature and features of the data, such that all teams are a priori satisfied that their approach would be suitable for the project. We further suggest that such a project uses artificial data generated by a third party. This data should be an artificial population with known parameters, but where individuals exhibit variations typical of real data. In such a scenario, teams would be blind to the parameters but have a real target to identify. Such an approach would further allow the assessment of statistical power, whereby some methods may be more sensitive than others and require smaller samples. It would also allow for quick replication studies, as well as evaluate the benefits (and potential pitfalls) of iterative analysis collaborations (e.g., Silberzahn et al., 2018). While such a project is more ambitious than the current one, it is an excellent complement to the first suggestion (multiple datasets, multiple methods) and a unique solution to generate best-practice guidelines.

It may also be worth pointing out that combining the many-analysts-approach with registered report format or results-masked-review format, as we used here, would require some further elaboration. Given that the introduction and methods cannot be modified after Stage 1 acceptance, it is difficult for lead authors to anticipate how exactly the collaboration will play out, and what steps of convergence may be taken. Perhaps these kind of multi-lab collaborations require a different submission format, one where the methods can be edited after all labs have submitted their results.

11.4. Conclusion

We suggest that, generally, researchers (whether individually or as part of collective efforts) plan for complementary if not competing analyses in their future projects, ideally in pre-registered studies where the benefits of this approach can be argued prior to the outcome of planned analyses. To this end, the OSF repository associated with the current paper, where readers may find code to reproduce the different analyses the teams have reported, may provide the research community with a useful set of tools to expand data exploration. While this paper focused on pupillometry, the methods that we have collectively made available could also be useful for other types of continuous data, such as heart-rate or EEG/ERPs, for example. In all cases, analyses (just like colleagues) may go further if they work together.

In the specific case of pupillometry, based on this project, we recommend that researchers assess that their findings are robust to changes in baseline parameters (and not, in the worst case, the direct outcome of arbitrary decisions about duration and location of baselines). We further recommend that methods that segment the temporal expression of pupil diameter into discrete temporal bins also assess their findings in relation to variations to the bin parameters (size, number, and position). More generally, we invite researchers interested in the minutiae of pupillometry analysis to join multi-lab collaborations as exciting opportunities to test and better understand these methods. In the case of data collection collaborations, early contributions at the design stage can help ensure that study design will be at the service of more than a single analysis method through consensus, allowing for comparisons between methods and finer-grained conclusions.

CRediT authorship contribution statement

Sylvain Sirois: Conceptualization, Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding aquisition. **Julie Brisson:** Methodology, Software, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Erik Blaser:** Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Giulia Calignano:** Data curation, Formal analysis, Software, Visualization, Writing - original draft, Writing - review & editing. **Jamie Donenfeld:** Writing – original draft, Writing – review & editing. **Robert Hepach:** Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Jean-Rémy Hochmann:** Software, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Zsuzsa Kaldy:** Writing – review & editing. **Ulf Liszkowski:** Writing – original draft, Writing – review & editing. **Marlena Mayer:** Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Shannon Ross-Sheehy:** Software, Validation, Formal analysis, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Sofia Russo:** Writing – review & editing. **Eloisa Valenza:** Writing – review & editing.

Data Availability

An OSF repository with raw, anonymized data, analysis code for each team, and video stimuli for the experiment can be found at https://osf.io/c5myh/.

Acknowledgments

We thank Martha Arterberry, Jordy Kaufmann, and one anonymous reviewer for comments on previous drafts. We also thank the infants who took part in this study, and their parents. This research was supported by funding from the Social Sciences and Humanities Research Council of Canada (grant 410-2010-2125) and research infrastructure support from the Canadian Foundation for Innovation to the first author. Each team in this collaboration contributed equally to the project..

Appendix

Chronology and Procedures of the Collaboration

This document provides a detailed and chronological overview of the key decisions that informed the production of the article *The pupil collaboration: A multi-lab, multi-method analysis of goal attribution in infants* (Sirois et al., 2023). Any further details may be requested from the corresponding author by email via sylvain.sirois@uqtr.ca.

February 2022

Infant Behavior and Development Editor Martha E. Arterberry and incomming member of the journal's Editorial Board Sylvain Sirois discuss potential new projects and special issues for the journal. One such idea is to invite multiple authors to analyze the same puillometry dataset, with a view of identifying best practice across variable methods that emerged within the last decade or so. Further exchanges explored how this could be instantiated.

March 2022

After exchanges between the Editor and the publisher, the initial format of the collaboration was:

1. Lead author provides a real, unpublished dataset
2. Lead author makes a public call for collaborators to analyze the dataset
3. Lead authors (as co-investigators behind the dataset) write introduction with focus on both the research question and the analysis collaboration
4. Lead authors write method section
5. Generate multiple results and short discussion sections for pupil analyses
6. A general discussion compares results and conclusions

The use of a real, unpublished dataset served two purposes. One, there would be no published results that could serve as a distracting benchmark for any team. Second, unlike artificial data that could have created effects to be discovered by appropriate analyses and thus create the possibility of winners and losers between collaborators, a real dataset provides some test of a research hypothesis without a known answer. A majority interpretation across teams would not, de facto, be correct. However, such an interpretation can highlight which aspects of varying methods share sensitivity to data features that uniquely informs future work.

The Editor proposes to serve as facilitator of the project and action editor for the ensuing paper. However, the Editor proposes that her role does not include duties that would constitute authorship.

April-May 2022

Lead author and Editor agree on the dataset to be used for the project. Lead author prepares data for sharing. Editor and lead author agree that only raw data files should be shared, without any of the data processing used by lead authors to protect the independence of findings between teams. They further agree that the usefulness of the project would be enhanced if, after all analyses are completed, a repository for the project with the raw data and analysis code from all teams were openly shared. Lead authors finalize introduction and method sections of planned paper.

June 2022

A decision is made to submit the manuscript for Results-Masked Review (RMR) and make the call for collaborators at the same time. Potential collaborators would be made aware that the manuscript is awaiting in-principle acceptance (IPA) prior to starting their work (so that they could choose to wait for IPA before committing time and resources to the project).

Editor and lead author agree that lead author will provide a list of all expressions of interest from would-be collaborators, and any justification for potential exclusion for Editor approval, to avoid any conflict of interest (real or apparent) for the lead author with this crucial task.

On June 9th, the manuscript was submitted for RMR with the journal. On the same day, a call for collaborators was posted on the listservs of the *International Congress of Infant Studies* (ICIS) and of the *Cognitive Development Society* (CDS), two international societies that reach a broad array of researchers interested in early development. The text of the call was the same in each of the two messages

and was as follows:

Invitation to collaborate on pupillometry analysis

Dear colleagues,

Infant Behavior and Development is trying a new article format, and we invite you to take part! We seek collaborators for a multi-lab, multi-method approach to pupillometry analysis. The study utilized the violation of expectation paradigm and data were collected using an eye-tracker, allowing for both looking time analyses and pupillometry analyses. Specifically, the study examined goal attribution using habituation then test trials, examining two independent-variables in a 2x2 repeated measures design. A single, documented dataset comprising raw data will be shared with collaborators. Armed with details about the method and the hypotheses of the study, you will be invited to submit your own results section and interim conclusions. You will also have the opportunity to collaborate on the general discussion, where your contribution is discussed in the context of the work by all collaborators. The lead authors of the paper provide the introduction and method, the dataset, and report on looking time and gaze data. Collaborations are sought exclusively for the analysis of pupil data.

Potential collaborators should

- Have a publication record with their proposed approach
- Be a team of no more than 3 people
- Be part of only one proposal
- Be able to submit their work by 15 September 2022

The aim of the collaboration is to evaluate whether and how different analytical approaches to pupillometry converge. Therefore, we aim to avoid repeat methods between collaborators/teams. As such

- The lead authors use functional data analysis, and so this approach is not invited
- Should two or more candidate collaborators propose the same approach, selection will be based on earliest publication with said approach.
- The authors and the Editor-in-Chief Martha Arterberry will confer on the final selection of collaborators.

Some additional details about the data

- The study design is a 2 x 2 repeated-measures approach
- Trials lasted ~12 s and pupils (stereo recording) were sampled @ 60 Hz
- There is at least 200 ms of pre-trial fixation data per trial, including pupils
- Training phase was infant-controlled (with a maximum of 20 trials)

Collaborators will be co-authors of the final paper, listed in alphabetical order after the two lead authors. The project is currently under review using the Results-Masked-Review format.

Finally, as the goal of the project is to identify best practice in the analysis of pupillometry, but also make these methods more accessible, it is expected that all collaborators will share their code (main script, bespoke functions, etc.) in a repository created for this project. This repository will be open-access after publication, so that any interested reader can explore the various methods by themselves. It is thus imperative that you agree to this aspect of the project prior to registering your interest.

Colleagues interested in taking part in this project should email Sylvain Sirois at sylvain.sirois@uqtr.ca with a description of the approach they would use, links to their publication(s) using this approach, and the names of up to two colleagues should they make a team proposal (only one member of a team should send a proposal). You can also email queries prior to sending a formal proposal. We would like to finalize the list of contributors by 15 July 2022.

Thank you in advance to those who will register interest.

Sylvain

July 2022

The Editor and lead author met on Zoom on 20 July to confirm invitations to colleagues who had expressed interest in taking part in the collaboration. We had initially received 7 expressions of interest, summarized in Table 1.

Table 1
Expression of interest with proposed methods, ranked by order in which they were received.

ID	Date	Method
1	10-Jun	Cluster mass permutation
2	10-Jun	PCA
3	10-Jun	Normalized time bins / multilevel modeling
4	10-Jun	Outcome-based linear mixed model
5	15-Jun	Multiverse / curve specification
6	10-Jul	Pre-processing/smoothing fixation contingent -> gaze-contingent linear mixed model
7	14-Jul	Critical-phase exclusion, interpolation, smoothing & baseline correction -> repeated mixed effect linear model

Of these, the second proposal (PCA) was withdrawn by the colleagues who proposed it when, after being provided more details about the data (specifically, the number of trials), they deemed their method would be unsuitable for this dataset as it requires

substantially more trials. The Editor and lead author agreed that all proposed methods would approach data analysis in a unique way (including planned pre-processing where specified), with no obvious repetition between them or with the lead authors' method. It was thus decided to invite the six remaining teams to the project, allowing for seven (including lead authors) distinct analysis approaches. The following email was sent on 21 July:

Dear colleagues,

I have now had the chance to consult with the journal's Editor, Martha Arterberry, and I am pleased to formally invite you to take part in this project. I would be very grateful if you could confirm your participation, especially in relation to the submission deadline outlined in the call for collaborators (i.e., 15 September 2022).

This somewhat impersonal email follows the approach to be favored moving forward, namely that the integrity of this project will be best served by radio silence between teams. Therefore, I will use an email list consisting of the lead (where applicable) member of each team, and BCC to everyone any and all updates about the project. We kindly ask, until submission, that you refrain from discussing your work on the project with colleagues outside of your team.

You should feel free to contact me about the project at any stage. We can exchange emails or have video calls. If anything from such exchanges appears significant for other groups, I plan to share such information with all teams. However, I will not share anything that pertains to individual teams' analyses (or identity). The goal is to ensure that all teams are on the same page prior to their individual efforts.

The manuscript is still under Results Masked Review, which we thought was the best approach for this project. This does mean that at this stage, we are still waiting for In Principle Acceptance. I will update you as soon as I can about the review outcome. You may obviously wish to start working on your analyses before the decision if this suits your planning better. This is entirely up to you. Depending on the review process (its timeline and outcome), we may have to move the deadline. However, as this is not a pre-registration and the data already exist from a specific method, the process should have no bearing on your analyses.

After I have received your participation confirmation, I will provide you next week with a link to the data and the paper (introduction and method).

Finally, as a minor aside, I will be on annual leave the first two weeks of August. I will make sure I have replied to any query on this project before I sign off. Likewise, I will prioritize this project upon returning.

Analyzers assemble!

Sylvain

As the email clarifies, the process moving forward is one where we stress no communication between teams until all analyses have been produced. The one, unavoidable exception to this rule concerns the lead author, who was known to all teams and would know the identity and methods of all teams. To protect the independence of the work, it was decided that all teams including lead authors would send their results to the Editor directly and exclusively, and that the Editor would send all sevensections back to the lead authors for the next stage of the project, with the constraint that no team would be able to alter their results thereafter.

On 27 July, the outcome of the RMR (invitation to resubmit after revisions) was sent to lead authors. The following email was sent to collaborators on the same day:

Dear collaborators,

Thank you for joining this exciting project. Here is some crucial information in order for you to proceed.

What we would like from you by 15 September (emailed to Martha Arterberry, marterbe@colby.edu)

A Results and Discussion Section (one section using the following numbered sections) complete with tables and figures:

3.0 Results and Discussion (add last names of team members here)

3.1 Pupillometry Method

3.2 Results

3.3 Discussion addressing specifically the hypotheses of the study and whether your results support them or not.

Please make notes (but do not put them in your discussion) regarding general comments about this collaboration for either the General Discussion and/or for the journal editor to keep in mind for future projects.

We also require supporting files that allow interested readers to reproduce your analyses (please comment your code!) or adopt your method, and any supplemental materials that explain your method(s) in details that go beyond the necessary information in section 3.1. Remember, we will have quite a few results sections! The main paper will focus on essential information. You can send us files in a.zip archive, or send us a link to a repository of your choice. Do not send or upload data to a repository, whether raw (we have them) or parsed/processed (which interested readers should be able to recreate from your supporting files). At the end of the project, we plan to have on

(*continued on next page*)

(*continued*)

OSF the raw data and the various documents required to reproduce any of the Results sections. Of course, if you use Matlab for example, all you need to provide are the scripts (and supporting functions that are not native) you used. Readers will have to have their own access to Matlab.

Access to files for the project

Files were uploaded to OSF, to which I provide you a read-only link.

OSF link: https://osf.io/c5myh/?view_only=b2c751424d61423188afe67a34cfbc23

If you have never used OSF, note that you can download individual files rather than opening them by clicking between the filename and date fields, which activates a download button. You can also download the whole contents of a folder to a zip file by clicking on the folder, which activates a download-as-zip button.

Please note that you can only share and use those files within your analysis team. Please do not distribute the files outside of your team, or allow them to be consulted in any way outside of your team.

In case you wonder: this is raw data. It has not been altered or beautified in any way. You may find that some trials or participants are not usable for your purposes. Just your typical infant data. As long as the usable n is reported in your analysis, and your procedure and/or code details selection/retention procedures, how you go about it is entirely your prerogative.

Sending your contributions

We wish that teams be blind to each other's work to protect the integrity of the project. As it were, I am also part of a team (i.e., lead authors). Therefore, we have decided that your work should not be sent to me, but to Martha Arterberry (marterbe@colby.edu), the journal Editor. She will only provide me with your work once she has received mine, which I will no longer be able to alter.

Compiling the final manuscript

We (lead authors) will compile the various results sections and draft a general conclusion. We will share that manuscript with all teams, and also put the draft general conclusion text online (e.g., google docs, privately) so that all teams can comment on the text and suggest modifications / additional text. When we are all agreed with the final text for the general discussion, we will send the full paper for final review.

Manuscript Update

Today, we received the reviews on the manuscript draft. Results-Masked-Review uses an in-principle acceptance process based on the introduction, method, and analytical plan. We are not there yet, but the comments suggest relatively straightforward changes. We are hopeful that we are close to the acceptance.

Final remarks

We are very excited about this project. We are also entering the Batcave stage, where you will all individually perform your magics in secrecy. So it may be a quiet period. However, you can contact me at any time with questions about the project. And if you raise a point of general interest, I will email everyone else about it so we can all work with the same information.

Good luck!

Sylvain

August 2022

Revised RMR manuscript submitted by lead authors on 18 August. IPA decision issued on 21 August.

October 2022

Editor shares with lead author the combined results sections for pupillometry analyses.

November 2022

The following email was sent to collaborators on 7 November:

Dear colleagues,

Thank you for your patience as I navigate a busy teaching semester. But thank you even more for your contribution to this project! I have read with much enthusiasm your R&D sections and am impatient to share the combined efforts with you. This email proposes a course of action for this very purpose.

1. By 18th November, we plan to share full, near-submission version of the final paper. This will combine the results-masked version, the looking time and gaze data, the 7 pupillometry sections, and a general discussion. This also involves substantial clerical work on figures, references etc.
2. The general discussion will include a section on the pupillometry results, and also a section of the collaboration per se. As you are co-authors of this paper, these are the two sections where we ask for your additional contribution. The paper will be made available in a Google doc file so that we can collectively work on those sections of text, which will be clearly identified. Your input is very welcomed.
3. We also asked that you review the CRediT roles we have assigned to you (and your team members) for your contribution. See https://www.elsevier.com/authors/policies-and-guidelines/credit-author-statement for additional details. If you think that anything should be removed or added, please let me know.
4. We would like the paper finalized for review by December 1st. So if you were able to see about points 2 and 3 by then, we'd be grateful

(*continued on next page*)

(*continued*)

5. If not by the 18th (priority will be the paper), then immediately after we will upload to the repository your materials (code, etc.) that you have provided to the Editor. Please double check your own folders (and subfolders, if any) at that time. Please let me know if you would like to add any additional files (readme, etc). A general, top-level readme file about the overall nature of the code, files, platforms/software, useful download links etc would be excellent.

The final version of the paper will then be submitted for review and, as such, should not be shared yet outside of your teams. For simplicity, we will work on the blind version of the paper. Here are the team codes, as we can waive anonymity (only within collaborators at this stage):

Team	**Members (corresponding author 1st)**
A	Blaser et al.
B	Calignano et al.
C	Hepach
D	Hochmann
E	Mayor & Liszkowski
F	Ross-Sheehy
G	Sirois & Brisson

Thanks again for fantastic work thus far!

Sylvain

On 21 November, the full paper was shared with collaborators on an online platform, with restricted access to collaborator email addresses. All teams were invited to edit and comment on the general discussion.

February 2023

The full paper was submitted for RMR - Stage 2 review on 28 February.

June 2023

Revise and resubmit decision on manuscript issued on 23 June. The following email was sent to collaborators:

Dear all,

The review is in! At first glance, this is neither terrible nor major. The worst of it is the deadline (7th of July!). Here is what I would plan to do:

- over the next week, see about making a first draft of the revision, along with a draft of the point-by-point response to the review
  - o if you have any thoughts/ideas/suggestions, feel free to email them to me
- when those drafts are ready, I would share them for collective revision before finalizing the resubmission

If you think you (or your team) may struggle with this tight and short timetable, please let me know so I can report to the editor.

Nearly there

Hope your summers are going well thus far

July 2023

On 21 July, the revised manuscript and draft response to comments were shared online (restricted to collaborators) by lead authors.

September 2023

After collaborative revision of the manuscript and related documents, submission of the revision to the journal.

References

Addyman, C., Rocha, S., & Mareschal, D. (2014). Mapping the origins of time: Scalar errors in infant time estimation. *Developmental Psychology, 50*(8), 2030–2035. https://doi.org/10.1037/a0037108

Anvari, F., & Lakens, D. (2018). The replicability crisis and public trust in psychological science. *Comprehensive Results in Social Psychology, 3*(3), 266–286. https://doi.org/10.1080/23743603.2019.1684822

Arnold, J.B. (2021). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. https://CRAN.R-project.org/package=ggthemes.

Aschersleben, G., Hofer, T., & Jovanovic, B. (2008). The link between infant attention to goal-directed action and later theory of mind abilities. *Developmental Science, 11*(6), 862–868. https://doi.org/10.1111/j.1467-7687.2008.00736.x

Aslin, R. N. (2007). What's in a look? *Developmental Science, 10*, 48–53. https://doi.org/10.1111/j.1467-7687.2007.00563.x

Bates, D.M. (2010). lme4: Mixed-effects modeling with R.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting Linear Mixed-Effects Models Usinglme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. Caccioppo, L. G. Tassinary, & G. Berntson (Eds.), *The Handbook of Psychophysiology*. Hillsdale, NJ: Cambridge University Press. https://doi.org/10.13140/2.1.2871.1369.

Behne, T., Carpenter, M., Call, J., & Tomasello, M. (2005). Unwilling Versus Unable: Infants' Understanding of Intentional Action. *Developmental Psychology, 41*(2), 328–337. https://doi.org/10.1037/0012-1649.41.2.328

Bellagamba, F., & Tomasello, M. (1999). Re-enacting intended acts: Comparing 12- and 18-month-olds. *Infant Behavior & Development, 22*, 227–282. https://doi.org/10.1016/S0163-6383(99)00002-8

Biro, S., & Leslie, A. M. (2007). Infants' perception of goaldirected action: development through cue-based bootstrapping. *Developmental Science, 10*, 379–398. https://doi.org/10.1111/j.1467-7687.2006.00544.x

Biro, S., Verschoor, S., & Coenen, L. (2011). Evidence for a unitary goal concept in 12-month-old infants. *Developmental Science, 14*, 1255–1260. https://doi.org/10.1111/j.1467-7687.2011.01042.x

Blaser, E., Eglington, L., Carter, A. S., & Kaldy, Z. (2014). Pupillometry reveals a mechanism for the Autism Spectrum Disorder (ASD) advantage in visual tasks. *Scientific Reports, 4*, 4301.

Bogartz, R. S., Shinskey, J. L., & Schilling, T. H. (2000). Object Permanence in Five-and-a-Half-Month-Old Infants. *Infancy, 1*(4), 403–428. https://doi.org/10.1207/S15327078IN0104_3

Bogartz, R. S., Shinskey, J. L., & Speaker, C. J. (1997). Interpreting infant looking: The event set × event set design. *Developmental Psychology, 33*(3), 408–422. https://doi.org/10.1037/0012-1649.33.3.408

Buresh, J. S., & Woodward, A. L. (2007). Infants track action goals within and across agents. *Cognition, 104*, 287–314. https://doi.org/10.1016/j.cognition.2006.07.001

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2021). Six solutions for more reliable infant research. *Infant and Child Development.* , Article e2296. https://doi.org/10.1002/icd.2296

Calignano, G., Girardi, P., & Altoé, G. (2023). First step into the pupillometry multiverse of developmental science. *Behavior Research Methods*. https://doi.org/10.3758/s13428-023-02172-8

Cannon, E., & Woodward, A. L. (2012). Infants generate goal-based action predictions. *Developmental Science, 15*, 292–298. https://doi.org/10.1111/j.1467-7687.2011.01127.x

Carpenter, M., Akhtar, N., & Tomasello, M. (1998). Fourteen- to 18-month-old infants differentially imitate intentional and accidental actions. *Infant Behavior and Development, 21*, 315–330. https://doi.org/10.1016/S0163-6383(98)90009-1

Chartier, C., Kline, M., McCarthy, R., Nuijten, M., Dunleavy, D. J., & Ledgerwood, A. (2018). The cooperative revolution is making psychological science better. *Association for Pychological Science: Observer, 31*(10).

Chen, Y.-C., & Westermann, G. (2018). Different novelties revealed by infants' pupillary responses. *Scientific Reports, 8*, 9533. https://doi.org/10.1038/s41598-018-27736-z

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

Choi, Y., Luo, Y., & Baillargeon, R. (2022). Can 5-month-old infants consider the perspective of a novel eyeless agent? New evidence for early mentalistic reasoning. *Child Development, 93*(2), 571–581. https://doi.org/10.1111/cdev.13707

Cohen, L. B. (2004). Uses and misuses of habituation and related preference paradigms. *Infant Child Development, 13*, 349–352. https://doi.org/10.1002/icd.355

Corkum, V., & Moore, C. (1998). The origins of joint visual attention in infants. *Dev Psychol, 34*, 28–38. https://doi.org/10.1037/0012-1649.34.1.28

Crivello, C., & Poulin-Dubois, D. (2018). Infants' false belief understanding: A non-replication of the helping task. *Cognitive Development, 46*, 51–57. https://doi.org/10.1016/j.cogdev.2017.10.003

Csibra, G., & Gergely, G. (1998). The teleological origins of mentalistic action explanations: A developmental hypothesis. *Developmental Science, 1*, 255–259. https://doi.org/10.1111/1467-7687.00039

Csibra, G. (2005). Mirror neurons and action observation: Is simulation involved. *ESF Interdisciplines*. ⟨https://eprints.bbk.ac.uk/id/eprint/29493⟩.

Csibra, G. (2008). Goal-attribution to inanimate agents by 6.5- month-old infants. *Cognition, 107*, 705–717. https://doi.org/10.1016/j.cognition.2007.08.001

Csibra, G. (2010). Recognizing communicative intentions in infancy. *Mind & Language, 25*, 141–168. https://doi.org/10.1111/j.1468-0017.2009.01384.x

Csibra, G., Gergely, G., Bíró, S., Koós, O., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of 'pure reason' in infancy. *Cognition, 72*, 237–267. https://doi.org/10.1016/s0010-0277(99)00039-6

Csink, V., Mareschal, D., & Gliga, T. (2021). Does surprise enhance infant memory? Assessing the impact of the encoding context on subsequent object recognition. *Infancy, 26*(2), 303–318. https://doi.org/10.1111/infa.12383

Davis-Kean, P. E., & Ellis, A. (2019). An overview of issues in infant and developmental research for the creation of robust and replicable science. *Infant Behavior and Development, 57*, Article 101339. https://doi.org/10.1016/j.infbeh.2019.101339

Deligianni, F., Senju, A., Gergely, G., & Csibra, G. (2011). Automated gaze-contingent objects elicit orientation following in 8-months-old infants. *Developmental Psychology, 47*, 1499–1503. https://doi.org/10.1037/a0025659

Dragicevic, P., Jansen, Y., Sarma, A., Kay, M., & Chevalier, F. (2019). Increasing the transparency of research papers with explorable multiverse analyses (May) *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 1–15.

Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., … Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology, 67*, 68–82. https://doi.org/10.1016/j.jesp.2015.10.012

Falck-Ytter, T., Gredebäck, G., & von Hofsten, C. (2006). Infants predict other people's action goals. *Nature Neuroscience, 9*(7), 878–879. https://doi.org/10.1038/nn1729

Fawcett, C., Arslan, M., Falck-Ytter, T., Roeyers, H., & Gredebäck, G. (2017). Human eyes with dilated pupils induce pupillary contagion in infants. *Scientific Reports, 7*(1), 1–7. https://doi.org/10.1038/s41598-017-08223-3

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., Hamlin, J. K., Hannon, E. E., Kline, M., Levelt, C., Lew-Williams, C., Nazzi, T., Panneton, R., Rabagliati, H., Soderstrom, M., Sullivan, J., Waxman, S., & Yurovsky, D. (2017). A Collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy, 22*(4), 421–435. https://doi.org/10.1111/infa.12182

Frank, M. C., Braginsky, M., Cachia, J., Coles, N., Hardwicke, T., Hawkins, R., … Williams, R. (2023). *Experimentology: An open science approach to experimental psychology methods*. Boston, MA: MIT Press.

Ganglmayer, K., Attig, M., Daum, M. M., & Paulus, M. (2019). Infants' perception of goal-directed actions: A multi-lab replication reveals that infants anticipate paths and not goals. *Infant Behavior and Development, 57*, Article 101340. https://doi.org/10.1016/j.infbeh.2019.101340

Geangu, E., Hauf, P., Bhardwaj, R., & Bentz, W. (2011). Infant pupil diameter changes in response to others' positive and negative emotions. *PLOS ONE, 6*(11), Article e27132. https://doi.org/10.1371/journal.pone.0027132

Gelman, A., & Loken, E. (2014). The statistical crisis in science: data-dependent analysis–a" garden of forking paths"–explains why many statistically significant comparisons don't hold up. *American Scientist, 102*(6), 460–466.

Geraci, A., Simion, F., & Surian, L. (2022). Infants' intention-based evaluations of distributive actions. *Journal of Experimental Child Psychology, 220*, Article 105429. https://doi.org/10.1016/j.jecp.2022.105429

Gergely, G., Nadasdy, Z., Csibra, G., & Biro, S. (1995). Taking the intentional stance at 12 months of age. *Cognition, 56*, 165–193. https://doi.org/10.1016/0010-0277(95)00661-h

Gerson, S., & Woodward, A. L. (2010). Building intentional action knowledge with one's hands. In S. P. Johnson (Ed.), *Neoconstructivism*. New York: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195331059.001.0001.

Gerson, S., & Woodward, A. L. (2014). Learning from their own actions: The unique effect of producing actions on infants' action understanding. *Child Development, 85*(*1*), 264–277. https://doi.org/10.1111/cdev.12115

Gredebäck, G., & Melinder, A. (2011). Teleological reasoning in 4-month-old infants: Pupil dilation and contextual constraints. *PLoS ONE, 6*(10), Article e26487. https://doi.org/10.1371/journal.pone.0026487

Green, A., Siposova, B., Kita, S., & Michael, J. (2021). Stopping at nothing: Two-year-olds differentiate between interrupted and abandoned goals. *Journal of Experimental Child Psychology, 209*, Article 105171. https://doi.org/10.1016/j.jecp.2021.105171
Gumbsch, C., Adam, M., Elsner, B., & Butz, M. V. (2021). Emergent goal-anticipatory gaze in infants via event-predictive learning and inference. *Cognitive Science, 45* (8), Article e13016. https://doi.org/10.1111/cogs.13016
Gustafsson, E., Brisson, J., Beaulieu, C., Mainville, M., Mailloux, D., & Sirois, S. (2015). How do infants recognize joint attention? *Infant Behavior and Development, 40*, 64–72. https://doi.org/10.1016/j.infbeh.2015.04.007
Gustafsson, E., Brisson, J., Mailloux, D., Mainville, M., Beaulieu, C., & Sirois, S. (2016). Do infants recognize engagement in social interactions? The case of face-to-face conversation. *Infancy, 21*(5), 685–696. https://doi.org/10.1111/infa.12135
Guttinger, S. (2020). The limits of replicability. *European Journal for Philosophy of Science, 10*(2), 10. https://doi.org/10.1007/s13194-019-0269-1
Haith, M. M. (1998). Who put the cog in infant cognition? Is rich interpretation too costly. *Infant Behavior & Development, 21*, 167–179. https://doi.org/10.1016/S0163-6383(98)90001-7
Hamlin, J. K., Hallinan, E. V., & Woodward, A. L. (2008). Do as I do: 7-month-old infants selectively reproduce others' goals. *Developmental Science, 11*, 487–494. https://doi.org/10.1111/j.1467-7687.2008.00694.x
Hellmer, K., Söderlund, H., & Gredebäck, G. (2018). The eye of the retriever: Developing episodic memory mechanisms in preverbal infants assessed through pupil dilation, 10-10 *Developmental Science, 21*(2). https://doi.org/10.1111/desc.12520.
Hepach, R. (forthcoming). Pupillometry in Developmental Psychology. In M. Papesh & S. Goldinger (Eds.), *Pupillometry: Cognition, Neuroscience, and Practical Applications*.
Hepach, R., Hedley, D., & Nuske, H. J. (2020). Prosocial attention in children with and without autism spectrum disorder: Dissociation between anticipatory gaze and internal arousal. *Journal of Abnormal Child Psychology, 48*(4), 589–605.
Hepach, R., Vaish, A., & Tomasello, M. (2012). Young children are intrinsically motivated to see others helped. *Psychological Science, 23*(9), 967–972.
Hepach, R., & Westermann, G. (2016). Pupillometry in infancy research. *Journal of Cognition and Development, 17*(3), 359–377. https://doi.org/10.1080/15248372.2015.1135801
Hernik, M., & Southgate, V. (2012). Nine-months-old infants do not need to know what the agent prefers in order to reason about its goals: On the role of preference and persistence in infants' goal-attribution. *Developmental Science, 15*(5), 714–722. https://doi.org/10.1111/j.1467-7687.2012.01151.x
Hochmann, J. R. (2022). Representations of abstract relations in infancy. *Open Mind, 6*, 291–310. https://doi.org/10.1162/opmi_a_00068
Hochmann, J. R., & Toro, J. M. (2021). Negative mental representations in infancy. *Cognition, 213*, Article 104599.
Hochmann, J.-R., & Papeo, L. (2014). The invariance problem in infancy: A pupillometry study. *Psychological Science, 25*(11), 2038–2046. https://doi.org/10.1177/0956797614547918
Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power. *The American Statistician, 55*(1), 19–24.
Hofer, T., Hohenberger, T., Hauf, P., & Aschersleben, G. (2008). The link between maternal interaction style and infant action understanding. *Infant Behavior and Development, 31*, 115–126. https://doi.org/10.1016/j.infbeh.2007.07.003
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
Jackson, I. R., & Sirois, S. (2022). But that's possible! Infants, pupils, and impossible events. *Infant Behavior & Development, 67*, Article 101710. https://doi.org/10.1016/j.infbeh.2022.101710
Jessen, S., Altvater-Mackensen, N., & Grossmann, T. (2016). Pupillary responses reveal infants' discrimination of facial emotions independent of conscious perception. *Cognition, 150*, 163–169. https://doi.org/10.1016/j.cognition.2016.02.010
Johnson, S., Slaughter, V., & Carey, S. (1998). Whose gaze will infants follow? Features that elicit gaze-following in 12-month-olds. *Developmental Science, 1*(2), 233–238. https://doi.org/10.1111/1467-7687.00036
Johnson, S. C., Booth, A., & O'Hearn, K. (2001). Inferring the goals of a nonhuman agent. *Cognitive Development, 16*, 637–656. https://doi.org/10.1016/S0885-2014(01)00043-0
Johnson, S. C., Shimizu, Y. A., & Ok, S.-J. (2007). Actors and actions: The role of agent behavior in infants' attribution of goals. *Cognitive Development, 22*(3), 310–322. https://doi.org/10.1016/j.cogdev.2007.01.002
Kagan, J. (2008). In defense of qualitative changes in development. *Child Development, 79*, 1606–1624. https://doi.org/10.1111/j.1467-8624.2008.01211.x
Kamewari, K., Kato, M., Kanda, T., Ishiguro, H., & Hiraki, K. (2005). Six-and-a-half-month-old children positively attribute goals to human action and to humanoid-robot motion. *Cognitive Development, 20*(2), 303–320. https://doi.org/10.1016/j.cogdev.2005.04.004
Karmiloff-Smith, A. (2012). Is development domain specific or domain general? A third alternative. In J. Shrager, & S. Carver (Eds.), *The journey from child to scientist: Integrating cognitive development and the education sciences* (pp. 127–140). Washington, DC US: American Psychological Association. https://doi.org/10.1037/13617-006.
Kiraly, I., Jovanovic, B., Prinz, W., Aschersleben, G., & Gergely, G. (2003). The early origins of goal-attribution in infancy. *Consciousness and Cognition, 12*, 752–769. https://doi.org/10.1016/S1053-8100(03)00084-9
Klein, R. A., Ratliff, K. A., Vianello, M., Adams, Bahník, Š., Bahník, Š., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology, 45*(3), 142–152. https://doi.org/10.1027/1864-9335/a000178
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software, 82*(13). https://doi.org/10.18637/jss.v082.i13
Lemche, E., Kreppner, J. M., Joraschky, P., & Klann-Delius, G. (2007). Attachment organization and the early development of internal state language: A longitudinal perspective. *International Journal of Behavioral Development, 31*(3), 252–262. https://doi.org/10.1177/0165025407076438
Leslie, A. M. (1994). Pretending and believing: issues in the theory of ToMM. *Cognition, 50*, 193–200. https://doi.org/10.1016/0010-0277(94)90029-9
Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology, 49*(4), 764–766.
Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *Journal of the American Statistical Association, 62*(318), 399–402.
López Pérez, D., Ramotowska, S., Malinowska-Korczak, A., Haman, M., & Tomalski, P. (2020). Working together to orient faster: The combined effects of alerting and orienting networks on pupillary responses at 8 months of age. *Developmental Cognitive Neuroscience, 42*. https://doi.org/10.1016/j.dcn.2020.100763
Lüdecke, D., & Lüdecke, M. D. (2015). Package 'sjPlot'. *R Package Version, 1*(9).
Luo, Y., & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science, 16*, 601–608. https://doi.org/10.1111/j.1467-9280.2005.01582.x
Luo, Y., & Beck, W. (2010). Do you see what I see? Infants' reasoning about others' incomplete perceptions. *Developmental Science, 13*, 134–142. https://doi.org/10.1111/j.1467-7687.2009.00863.x
Luo, Y. (2011). Three-month-old infants attribute goals to a non-human agent. *Developmental Science, 14*, 453–460. https://doi.org/10.1111/j.1467-7687.2010.00995.x
Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods, 164*(1), 177–190.
Mathôt, S. (2018). Pupillometry: Psychology, physiology, and function. *Journal of Cognition, 1*(1). 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.
Mathôt, S., van der Linden, L., Grainger, J., & Vitu, F. (2015). The pupillary light response reflects eye-movement preparation. *Journal of Experimental Psychology Human Perception and Performance, 41*(1), 28–35.
Mathôt, S., & Vilotijević, A. (2023). Methods in cognitive pupillometry: Design, preprocessing, and statistical analysis. *Behavior Research Methods, 55*(6), 3055–3077. https://doi.org/10.3758/s13428-022-01957-7

Meltzoff, A. N. (1995). Understanding the intention of others: reenactment of intended acts by 18-month-old children. *Developmental Psychology, 31*, 838–850. https://doi.org/10.1037/0012-1649.31.5.838

Moll, H., Koring, C., Carpenter, M., & Tomasello, M. (2006). Infants determine others' focus of attention by pragmatics and exclusion. *Journal of Cognition and Development, 7*(3), 411–430. https://doi.org/10.1207/s15327647jcd0703_9

Morita, T., Slaughter, V., Katayama, N., Kitazaki, M., Kakigi, R., & Itakura, S. (2012). Infant and adult perceptions of possible and impossible body movements: an eyetracking study. *Journal of Experimental Child Psychology, 113*(3), 401–414. https://doi.org/10.1016/j.jecp.2012.07.003

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Pätzold, W., & Liszkowski, U. (2020). Pupillometric VoE paradigm reveals that 18- but not 10-month-olds spontaneously represent occluded objects (but not empty sets. *PLoS One, 15*(4), Article e0230913. https://doi.org/10.1371/journal.pone.0230913

Perner, J., & Ruffman, T. (2005). Infants' insight into the mind: How deep. *Science, 308*, 214–216. https://doi.org/10.1126/science.1111656

Phillips, A. T., Wellman, H. M., & Spelke, E. S. (2002). Infants' ability to connect gaze and emotional expression to intentional action. *Cognition, 85*, 53–78. https://doi.org/10.1016/s0010-0277(02)00073-2

Poulin-Dubois, D., Rakoczy, H., Burnside, K., Crivello, C., Dörrenberg, S., Edwards, K., Krist, H., Kulke, L., Liszkowski, U., Low, J., Perner, J., Powell, L., Priewasser, B., Rafetseder, E., & Ruffman, T. (2018). Do infants understand false beliefs? We don't know yet – A commentary on Baillargeon, Buttelmann and Southgate's commentary. *Cognitive Development, 48*, 302–315. https://doi.org/10.1016/j.cogdev.2018.09.005

Quartz, S., & Sejnowski, T. (1997). The neural basis of cognitive development: A constructivist manifesto. *Behavioral and Brain Sciences, 20*(4), 537–556. https://doi.org/10.1017/S0140525X97001581

Quillien, T., & German, T. C. (2021). A simple definition of 'intentionally'. *Cognition, 214*, Article 104806. https://doi.org/10.1016/j.cognition.2021.104806

Ramsay, J., & Silverman, B. (1997). *Functional Data Analysis*. New York: Springer. https://doi.org/10.1007/978-1-4757-7107-7

Reid, V. M., Csibra, G., Belsky, J., & Johnson, M. H. (2007). Neural correlates of the perception of goal-directed action in infants. *Acta Psychologica, 124*, 129–138. https://doi.org/10.1016/j.actpsy.2006.09.010

Reiss, R. D., Thomas, M., & Reiss, R. D. (1997). *Statistical analysis of extreme values* (Vol. 2). Basel: Birkhäuser,.

Romero, F. (2019). Philosophy of science and the replicability crisis. *Philosophy Compass, 14*(11), Article e12633. https://doi.org/10.1111/phc3.12633

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

Rubio-Fernández, P. (2019). Publication standards in infancy research: Three ways to make Violation-of-Expectation studies more reliable. *Infant Behavior & Development, 54*, 177–188. https://doi.org/10.1016/j.infbeh.2018.09.009

Saylor, M. M., & Ganea, P. A. (2007). Infants interpret ambiguous requests for absent objects. *Developmental Psychology, 43*(3), 696–704. https://doi.org/10.1037/0012-1649.43.3.696

Schlottmann, A., & Ray, E. (2010). Goal-attribution to schematic animals: do 6-month-olds perceive biological motion as animate? *Developmental Science, 13*, 1–10. https://doi.org/10.1111/j.1467-7687.2009.00854.x

Schweinsberg, M., Feldman, M., Staub, N., van den Akker, O. R., van Aert, R. C. M., van Assen, M. A. L. M., Liu, Y., Althoff, T., Heer, J., Kale, A., Mohamed, Z., Amireh, H., Venkatesh Prasad, V., Bernstein, A., Robinson, E., Snellman, K., Amy Sommer, S., Otner, S. M. G., Robinson, D., & Luis Uhlmann, E. (2021). Same data, different conclusions: Radical dispersion in empirical results when independent analysts operationalize and test the same hypothesis. *Organizational Behavior and Human Decision Processes, 165*, 228–249. https://doi.org/10.1016/j.obhdp.2021.02.003

Serra-Garcia, M., & Gneezy, U. (2021). Nonreplicable publications are cited more than replicable ones. *Science Advances, 7*(21), eabd1705. https://doi.org/10.1126/sciadv.abd1705

Shimizu, Y. A., & Johnson, S. C. (2004). Infants' attribution of a goal to a morphologically novel agent. *Developmental Science, 7*, 425–430. https://doi.org/10.1111/j.1467-7687.2004.00362.x

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahník, Š., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Dalla Rosa, A., Dam, L., Evans, M. H., Flores Cervantes, I., & Nosek, B. A. (2018). Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science, 1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*(11), 1359–1366.

Singmann, H., & Kellen, D. (2019). An Introduction to Mixed Models for Experimental Psychology. In D. H. Spieler, & E. Schumacher (Eds.), *New Methods in Cognitive Psychology* (pp. 4–31). Psychology Press.

Sirois, S. (2022). The seventh solution: A commentary on Byers-Heinlein, Bergmann, and Savalei (2021). *Infant and Child Development.* , Article e2351. https://doi.org/10.1002/icd.2351

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

Sirois, S., & Jackson, I. (2007). Social cognition in infancy: a critical review of research on higher-order abilities. *European Journal of Developmental Psychology, 4*(1), 46–64. https://doi.org/10.1080/17405620601047053

Sirois, S., & Jackson, I. R. (2012). Pupil Dilation and Object Permanence in Infants. *Infancy, 17*, 61–78. https://doi.org/10.1111/j.1532-7078.2011.00096.x

Sodian, B., Schoeppner, B., & Metz, U. (2004). Do infants apply the principle of rational action to human agents. *Infant Behavior and Development, 27*, 31–41. https://doi.org/10.1016/j.infbeh.2003.05.006

Sommerville, J. A., Hildebrand, E. A., & Crane, C. C. (2008). Experience matters: The impact of doing versus watching on infants' subsequent perception of tool use events. *Developmental Psychology, 44*, 1249–1256. https://doi.org/10.1037/a0012296

Sommerville, J. A., Woodward, A. L., & Needham, A. (2005). Action experience alters 3-month-old infants' perception of others'actions. *Cognition, 96*, B1–B11. https://doi.org/10.1016/j.cognition.2004.07.004

Southgate, V., Johnson, M. H., & Csibra, G. (2008). Infants attribute goals even to biomechanically impossible actions. *Cognition, 107*(3), 1059–1069. https://doi.org/10.1016/j.cognition.2007.10.002

Steegen, S., Tuerlinckx, F., Gelman, A., & Vanpaemel, W. (2016). Increasing transparency through a 9 multiverse analysis. *Perspectives on Psychological Science, 11*(5), 702–712. https://doi.org/10.1177/1745691616658637

Team, R. C. (2020). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing,.

The ManyBabies Consortium. (2020). Quantifying Sources of Variability in Infancy Research Using the Infant-Directed-Speech Preference. *Advances in Methods and Practices in Psychological Science, 3*(1), 24–52. https://doi.org/10.1177/2515245919900809

Ting, F., He, Z., & Baillargeon, R. (2021). Five-month-old infants attribute inferences based on general knowledge to agents. *Journal of Experimental Child Psychology, 208*, Article 105126. https://doi.org/10.1016/j.jecp.2021.105126

Tomasello, M., & Rakoczy, H. (2003). What makes human cognition unique? From individual to shared to collective intentionality. *Mind & Language, 18*(2), 121–147. https://doi.org/10.1111/1468-0017.00217

Tomasello, M., Carpenter, M., Call, J., Behne, T., & Moll, H. (2005). Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences, 28*, 675–735. https://doi.org/10.1017/S0140525X05000129

Upshaw, M. B., Kaiser, C. R., & Sommerville, J. A. (2015). Parents' empathic perspective taking and altruistic behavior predicts infants' arousal to others' emotions. *Frontiers in Psychology, 6*. https://doi.org/10.3389/fpsyg.2015.00360

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

Verschoor, S. A., Paulus, M., Spapė, M., Biro, S., & Hommel, B. (2015). The developing cognitive substrate of sequential action control in 9- to 12-month-olds: Evidence for concurrent activation models. *Cognition, 138C*, 64–78. https://doi.org/10.1016/j.cognition.2015.01.005

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832.

Wickham, H., Chang, W., & Wickham, M. H. (2016). Package 'ggplot2'. Create elegant data visualisations using the grammar of graphics. *Version, 2*(1), 1–189.

Woodward, A. L., & Guajardo, J. J. (2002). Infants' understanding of the point gesture as an object-directed action. *Cognitive Development, 17*, 1061–1084. https://doi.org/10.1016/S0885-2014(02)00074-6

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition, 69*, 1–34. https://doi.org/10.1016/S0010-0277(98)00058-4

Woodward, A. L. (2003). Infants' developing understanding of the link between looker and object. *Developmental Science, 6*(3), 297–311. https://doi.org/10.1111/1467-7687.00286

Woodward, A. L. (2009). Infants' grasp of others' intentions. *Current Directions in Psychological Science, 18*, 53–57. https://doi.org/10.1111/j.1467-8721.2009.01605.x

Woodward, A. L. (2005). The infant origins of intentional understanding. In R. V. Kail (Ed.), *Advances in child development and behavior* (Vol.33, pp. 229–262). Oxford: Elsevier.

Woodward, A. L., Sommerville, J. A., & Guajardo, J. J. (2001). How infants make sense of intentional action. In B. Malle, L. Moses, & D. Baldwin (Eds.), *Intentions and intentionality: Foundations of Social Cognition* (pp. 149–169). Cambridge, MA: MIT Press.

Yoon, J. M. D., Johnson, M. H., & Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences, 105*(36), 13690–13695. https://doi.org/10.1073/pnas.0804388105

Yu, C., Yurovsky, D., & Xu, C. (2012). Visual data mining: An exploratory approach to analyzing temporal patterns of eye movements. *Infancy, 17*(1), 33–60. https://doi.org/10.1111/j.1532-7078.2011.00095.x

Zhang, F., Jaffe-Dax, S., Wilson, R. C., & Emberson, L. L. (2018). Prediction in infants and adults: A pupillometry study. *Developmental Science, 22*(4), 1–9. https://doi.org/10.1111/desc.12780