# Does the understanding of complex dynamic events at 10 months predict vocabulary development?*

SAMANTHA DURRANT* iD

*University of Liverpool*

ANDREW JESSOP*

*Max Planck Institute for Psycholinguistics*

FRANKLIN CHANG

*Kobe City University for Foreign Studies*

AMY BIDGOOD iD

*University of Salford*

MICHELLE S. PETER

*University of Liverpool*

JULIAN M. PINE

*University of Liverpool*

AND

CAROLINE F. ROWLAND

*University of Liverpool, Max Planck Institute for Psycholinguistics, and Donders Institute for Brain, Cognition and Behaviour*

---

1

ABSTRACT

By the end of their first year, infants can interpret many different types of complex dynamic visual events, such as caused-motion, chasing, and goal-directed action. Infants of this age are also in the early stages of vocabulary development, producing their first words at around 12 months. The present work examined whether there are meaningful individual differences in infants' ability to represent dynamic causal events in visual scenes, and whether these differences influence vocabulary development. As part of the longitudinal Language 0–5 Project, 78 10-month-old infants were tested on their ability to interpret three dynamic motion events, involving (a) caused-motion, (b) chasing behaviour, and (c) goal-directed movement. Planned analyses found that infants showed evidence of understanding the first two event types, but not the third. Looking behaviour in each task was not meaningfully related to vocabulary development, nor were there any correlations between the tasks. The results of additional exploratory analyses and simulations suggested that the infants' understanding of each event may not be predictive of their vocabulary development, and that looking times in these tasks may not be reliably capturing any meaningful individual differences in their knowledge. This raises questions about how to convert experimental group designs to individual differences measures, and how to interpret infant looking time behaviour.

KEYWORDS: causal events, infant perception, vocabulary development, individual differences

## 1. Introduction

There are substantial individual differences in the age and rate at which children acquire vocabulary. These emerge early and increase in magnitude over time (e.g., Fenson et al., 1994). Evidence from the Stanford Wordbank (http://wordbank.stanford.edu) suggests that this variance in vocabulary acquisition is a cross-linguistic phenomenon (Frank, Braginsky, Yurovsky, & Marchman, 2017); in 23 different languages, 12-month-olds on the 80th percentile produce an average of 13 words, whereas those on the 20th percentile produce around 1.23 words. Although many of the children with the smallest vocabulary sizes at this age will ultimately be diagnosed with a developmental language disorder, there remain substantial and meaningful developmental differences among children in the typical range. Understanding the causes of these differences is a crucial step towards building comprehensive models of language acquisition. Previous studies of typically developing children have considered whether individual differences in vocabulary development are the

2

result of variance in the children's environment (e.g., parental input: Hoff, 2006) and cognitive abilities (e.g., speed of linguistic processing: Fernald, Perfors, & Marchman, 2006; phonological working memory: Gathercole & Baddeley, 1989; and statistical learning: Kidd, Junge, Spokes, Morrison, & Cutler, 2018). The aim of the present study was to examine whether the capacity to understand complex dynamic events in the world is also a contributing factor by testing these skills at the earliest stages of vocabulary development. That is, at 9 to 10 months old, do individual differences in representations of the conceptual world form part of the explanation for the variance observed in vocabulary development?

Since the relationship between words and their referents can be seen as largely arbitrary (Saussure, 1916), it is often assumed that word meaning is derived from the visual–spatial understanding of events. Mandler (1992) argued that perceptual analysis is critical in developing meaning representations that support later language learning. Part of this analysis involves abstraction from the low-level visual input into higher-level image schemas: complex relational representations of events (see Figure 1 for an example of caused motion), which according to cognitive linguistic accounts of linguistic knowledge are central to language representation (Langacker, 1987; Talmy, 1988) and language acquisition (Tomasello, 2003). Although there is variation across languages in the visual–spatial components that are expressed in language (e.g., Korean verbs distinguish tight/loose fit; Choi & Bowerman, 1991), many of the visual features that can be expressed in language are present universally before language acquisition begins (Chang, Choi, & Ko, 2015; Hespos & Spelke, 2004). Therefore, speakers' shared understanding of language meaning must be based on some shared non-linguistic understanding of events, and visual–spatial information appears to be the most likely source.

In this study, we test the idea that, if language meaning depends on the ability to understand events perceptually, then variation in vocabulary acquisition may depend on variation in visual–spatial event understanding. Since visual perception is a complex system with many subcomponents (Marr, 1982), it is necessary to restrict our examination of this relationship to a small set of features. To do this, we selected three aspects of event understanding that
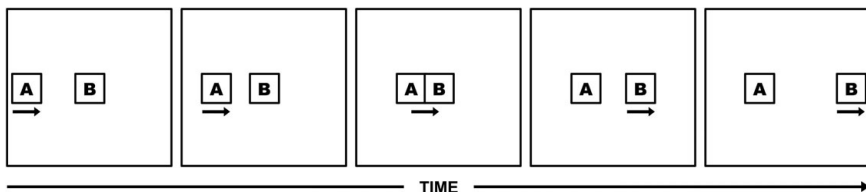


Fig. 1 An illustration of Michotte's (1946) launching display.

3

have been tested in infants: causality, durational actions, and goal-directed actions. These components were selected to be as abstract as possible, such that they could potentially apply to a wide range of words.

The first feature we examined was causality between two entities, which is critical to distinguish verbs that describe self-generated motion and those that describe externally caused motion (e.g., "the car moved" vs. "he pushed the car"). Nouns are also defined in terms of causal features (e.g., a match is a stick that causes fire when contact is made with another object). Causality is a very general concept that applies across a range of events, but it is most easily seen in scenes of pushing. Research on causality in these types of scenes was pioneered by Michotte (1946), who tested adult participants on their perception of visual scenes using *launching* displays illustrated in Figure 1. In these scenes, shape A moved directly towards shape B, which was stationary. When shape A made physical contact with shape B, shape A stopped moving and shape B immediately moved away along the same trajectory. Viewers reliably interpreted these sequences as causal events in which shape A *launches* shape B into motion, akin to one billiard ball hitting another, even though these were animated static images without any causal link in the motion (Hume, 1748). Similar launching event studies have found that infants can discriminate causal from non-causal events (Leslie, 1982, 1984; Leslie & Keeble, 1987; Oakes, 1994; Oakes & Cohen, 1990), but that the impression of causality is reduced in adults and children when there is no contact between the objects or when there is a delay in the launching of the second object. For example, Oakes and Cohen (1990) habituated 6- and 10-month-olds to either a causal launch display (Figure 1) or a non-causal delayed launching or no collision display. They were then shown novel non-causal or causal displays, respectively. The 10-month-olds, but not the 6-month-olds, showed a significant increase in looking to the event type that differed in causality from the event seen during habituation. This suggests that, by 10 months, infants can distinguish between causal and non-causal events based on only spatial and temporal characteristics. Importantly, these looking patterns were not attributable to the novelty of the alternative event; when habituated to a non-causal event, the infants dishabituated to a novel causal event but not to a novel non-causal event. According to the authors, this showed that the infants had identified the similarity in causality status between the familiar and novel non-causal events. Other studies have shown that infants as young as 6 months make similar distinctions (Cohen & Amsel, 1998; Leslie, 1984), though, in this case, performance was sensitive to the specific objects involved in the events (e.g., changing the agent and patient between trials affected performance; Cohen & Oakes, 1993). Overall, this work suggests that infants can already understand some of the features used to identify agents and patients in causal events at the very early stages of language acquisition.

4

Causal pushing events focus on the brief interaction between objects around contact, but many events require that children aggregate information over a period of time. For example, walking and running can only be distinguished by observing the action over time (running leads to more movement). Likewise, some noun-referents, like snack and meal, can be distinguished by the way that they are performed over time (a meal involves eating over a longer period of time). Thus, another ability that is important for understanding word meaning is the ability to track objects' interactions over time. This ability is tapped in studies of chasing events (Gao, Newman, & Scholl, 2009; Gao & Scholl, 2011), where participants see identical objects on the screen and must identify the chaser in the event. Participants do this by tracking the angle of motion of all the objects relative to the other objects (the chaser moves towards the chasee). Frankenhuis, House, Barrett, and Johnson (2013) have shown that 4- and 10-month-old infants can distinguish a chasing relationship from a scene with random motion. By manipulating different aspects of the chasing motion, the authors showed that the preference for chasing was primarily due to two visual cues; the acceleration of the patient to "escape" the agent, and the "heat-seeking" attraction of the agent to the patient. This is consistent with the findings of other infant studies using an eye-tracking paradigm (Galazka & Nyström, 2016), and experiments with older children or adult participants (Barrett, Todd, Miller, & Blythe, 2005; Gao et al., 2009; Gao & Scholl, 2011; Meyerhoff, Papenmeier, Jahn, & Huff, 2013).

Third, and finally, many word meanings are distinguished by observable physical changes. For example, children come to understand that the word "try" refers to intending to do something, while the phrase "by accident" is used when an action was unintentional (Astington, 1986). Work on child social cognition has found that infants can also use motion cues to infer the intentions and goals motivating an agent's behaviour (e.g., Csibra, 2008; Csibra, Gergely, Bıŕó, Koós, & Brockbank, 1999; Gergely, Nádasdy, Csibra, & Bíró, 1995; Luo, 2011; Luo & Baillargeon, 2005; Shimizu & Johnson, 2004; Woodward, 1998). Luo (2011) examined whether 3-month-old infants can attribute goals and preferences to objects that appeared to be inanimate. The infants saw physical object displays containing a moving box positioned between a cone and a cylinder. They were familiarised to a scene in which the box repeatedly approached one of the objects, before being tested on trials where the box either approached the alternative object or continued to approach the original object. It was found that infants watched for longer when the box approached the new object than the old object, even when the location of the target objects was switched (i.e., when the direction of movement towards the new object matched the familiarisation trials). Luo suggested that the infants see the box as having a preference for the old object and hence they were surprised to see it switch goals. Other research has similarly found that infants will attribute goals

5

and intentions to simple shapes (for a review, see Baillargeon, Scott, & Bian, 2016). Collectively, these studies suggest that the critical feature that triggers the perception of agency is evidence of internal control, such as when a self-propelled agent consistently chooses to approach a goal object, even after it has changed position in the display. Infants also appear to make goal attributions based on an expectation that agents will act rationally; for example, taking the shortest available path to reach the goal object (e.g., Csibra et al., 1999; Gergely et al., 1995).

In sum, infants appear to be capable of using spatial and motion cues to identify physical causality by 10 months (Oakes & Cohen, 1990), intentional chasing by 4 months (Frankenhuis et al., 2013), and goal-directed behaviour by 3 months (Luo, 2011). However, it is unclear whether there are meaningful individual differences in these abilities; that is, whether some children have a deeper understanding of complex dynamic events or are faster at processing such visual scenes compared to others of the same age. There is reason to think that this may be the case since robust individual differences with developmental implications have been observed in a range of cognitive domains. For example, rate of habituation (i.e., speed of decrement of attention to a repeatedly presented stimulus) and recognition memory (i.e., ability to discriminate between familiar and unfamiliar stimuli) in infants younger than 1 year have been reported, and even shown to meaningfully predict IQ in later childhood (e.g., Bornstein & Sigman, 1986; Kavšek, 2004; McCall & Carriger, 1993). Yet, Gampe, Keitel, and Daum (2015) observed a lack of individual consistency in infants' anticipatory looking times when they were regularly tested between the ages of 7 and 12 months with an action-perception task where an agent reaches for a ball and drops it into a bucket (see also Melzer, Prinz, & Daum, 2012). Thus, the first aim of the present work was to test whether there are meaningful individual differences in the ability of 9- to 10-month-olds to interpret caused-motion (Oakes & Cohen, 1990), chasing interactions (Frankenhuis et al., 2013), and goal-directed behaviour (Luo, 2011). Since all three tasks tap into some aspect of the child's conceptual understanding, it is reasonable to assume that we would expect to find correlations across the three tasks; for example, that children with sophisticated conceptual understanding in one task (e.g., discriminating causal from non-causal motion events) should also demonstrate sophisticated conceptual understanding in another (e.g., discriminating goal-directed from non-goal-directed events).

The three tasks used in the present research used habituation/familiarisation (in the case of caused-motion and goal-directed behaviour) and preferential looking (in the case of chasing) designs. Extracting measures of individual differences in these three tasks presents a methodological challenge. In habituation/familiarisation designs, children are habituated to one stimulus (i.e., a stimulus is repeatedly presented), before being shown a novel stimulus. In

6

preferential looking designs, children are shown two events simultaneously (e.g., a chasing event vs. random movement). In both designs, the outcome measure is a looking-time difference analysed at the group level; if the infants, as a group, tend to look for longer at one event than the other, all infants at that age are credited with the relevant ability. However, to test for individual differences, we need to assess the outcome of these tasks at the participant level. In this study, we used looking-time differences, relying on the assumption that individual looking-time differences are meaningful indicators of an infant's ability. This assumption is validated by previous research, which has observed robust relationships between vocabulary knowledge and looking times across a range of tasks, including those assessing lexical speed of processing (Fernald et al., 2006; Lany, 2018), recognition memory (Gathercole & Baddeley, 1989), statistical learning (Kidd et al., 2018), word segmentation (Singh, Steven Reznick, & Xuehua, 2012), sentence processing (Mani & Huettig, 2012), and visual sequence learning (Shafto, Conway, Field, & Houston, 2012). Note too, that the use of parametric tests to analyse group behaviour in infant looking-time studies also assumes that individual differences in looking times directly measure performance. For example, in the causal motion study of Oakes and Cohen (1990), the statistical tests used rely on the assumption that a child who dishabituates more (i.e., looks substantially longer at the novel than the habituated event) not only knows more about causal motion than a child who does not dishabituate at all, but also knows more than a child who only dishabituates a little (i.e., only looks a little longer at the novel event). Thus, for present purposes, looking-time differences are treated as direct measurements of infants' event understanding (to foreshadow the results, the findings suggest that this may not be a legitimate assumption; the implications of this for the interpretation of looking-time measures more generally are considered further in the 'General discussion').

The second aim of the present research was to determine whether infants' ability to interpret complex events influences their ability to learn new words. All theories of language acquisition assume that visual–spatial event understanding is an important part of language acquisition (Goldin-Meadow & Mylander, 1998; Landau & Jackendoff, 1993; Mandler, 1992; Tomasello & Vaish, 2013). To date, the nature of this relationship has not been tested. In this work, we examined whether individual differences in looking times on our three event perception tasks at 9–10 months predicted overall vocabulary size, and action verb vocabulary size, both concurrently and longitudinally.

In sum, the goal of this study was to determine whether there are individual differences in infants' ability to represent dynamic causal events in visual scenes and whether these differences contribute to individual differences in vocabulary size. Infants aged between 9 and 10 months took part in three studies designed to investigate their ability to interpret dynamic causal events;

a caused-motion event (Oakes & Cohen, 1990), a chasing event (Frankenhuis et al., 2013), and a goal-directed event (Luo, 2011). Infants were tested at 9–10 months, since this is the point at which children are starting to learn to comprehend, if not produce, words. Thus, their ability to learn may be influenced by their conceptual representations. By the age tested, above-chance performance at the group level is observed in all three of the original versions of these tasks. Whilst 9–10 months is an older age than was tested in the original versions of two of these tasks (goal-directed behaviour and intentional chasing), there is no empirical reason to assume that there will be no meaningful variation in their understanding at a later time-point. First, we assessed whether the group effects reported by the original authors could be replicated. Second, we tested whether looking-time behaviour in these tasks provided reliable, and correlated, measures of individual differences in ability. Finally, we examined whether individual differences correlated with concurrent and later vocabulary size. The results of our planned analyses are dependent on two critical assumptions about how event knowledge can be measured in infants. First, it is commonly, and implicitly, assumed that looking times linearly reflect knowledge, such that longer looking times are indicative of a deeper understanding of the stimulus. The second assumption is that the type of looking preference (novelty or familiarity) is meaningful. In a series of exploratory analyses and simulations, we tested the validity of these assumptions.

## 2. Method

We tested whether 10-month-old infants could interpret dynamic events in three tasks requiring the attribution of caused-motion, chasing, and goal-directed action. Data on vocabulary size were collected at multiple time-points using parent-completed questionnaires and the full Bayley Scales of Infant and Toddler Development, Third Edition (BSID-III; Bayley, 2006) was administered at 16–17 months.

### 2.1. PARTICIPANTS

These studies form part of a larger longitudinal project in the northwest of England called the Language 0–5 Project. Ninety-five typically developing monolingual participants were recruited to take part at the outset of the project (see <https://osf.io/wdy7c/>). The three experimental studies presented here were completed in a single session when the participants were aged between 9 and 10 months. A total of 78 infants (41 females) attended this session (mean age = 9;24 months, range = 9;12–10;18).

8

## 2.2. MATERIALS AND PROCEDURE

For all three studies, infants were seated either in a high-chair (with their caregiver to their left facing away from the screen) or on their parents' lap 60cm away from a widescreen TV (3846 × 1080). During testing, infants were allowed a dummy/pacifier or snack if required. Studies were presented using the Look program (Meints, Plunkett, & Harris, 2008), and controlled by an experimenter monitoring the infant's behaviour via a webcam (see Figure 2 for set-up). In all tasks, trials were initiated by the experimenter when the infant was looking towards the screen, although there were differences in the durations the stimuli were presented for and in how trials were terminated (further details are provided in the next section). These differences between tasks were maintained to adhere to the original designs as closely as possible. As we were testing older children than in the original studies and on multiple tasks in a single session, we aimed to replicate every task as closely as possible so as to minimise the reasons for any failure to replicate. Each task was presented twice
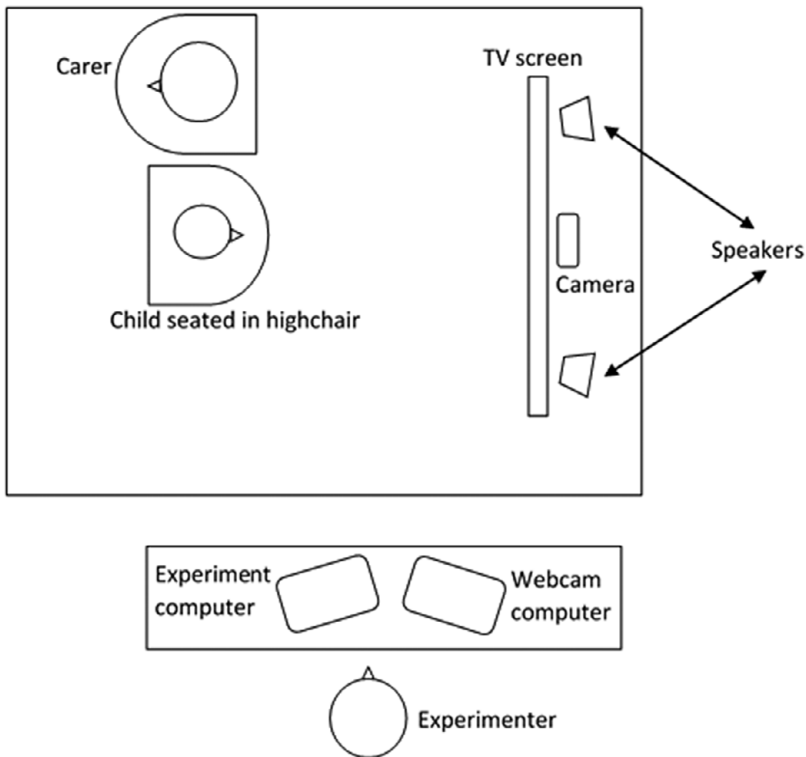


Fig. 2 Diagram of the laboratory set-up.

9

during the session, with presentation order organised into three pairings. All infants completed the tasks in the same order, with each task presented both in the first and second position within a pairing. Between pairings, there was a short play break where infants left the testing booth and played with the experimenter and their caregiver. The duration of the break between pairings varied for each infant ($M = 10{:}22$ minutes, range $= 1{:}43$–$22{:}10$ minutes). The structure of the testing session was as follows:

1. Task 1a and Task 2a
2. Play break
3. Task 3a and Task 1b
4. Play break
5. Task 2b and Task 3b

### 2.3. TASK 1: DISTINGUISHING CAUSAL FROM NON-CAUSAL MOTION

The first task was an adaptation of Oakes and Cohen (1990) used to assess infants' ability to distinguish between causal and non-causal events with novel inanimate objects. Oakes and Cohen presented infants aged 6 and 10 months with four different scenes of real-world objects interacting in different ways: two of the videos demonstrated two different non-causal events, one a casual event, and one a single object control event. There were two phases to each testing session, a habituation phase and a test phase. In the habituation phase, infants were shown a maximum of 20 trials displaying either the causal event or one of the non-causal events. Individual trials terminated if the infant looked away for 500 milliseconds (ms), and the habituation phase ended when the infants' looking time in a block of 4 trials reduced below 50% of the looking time in the first four trials. Following habituation, all four events were presented, starting with the habituated event and ending with the control event. The infants showed an increase in looking to the event type they were not habituated to. Most importantly for our study, when habituated to a non-causal event, the infants increased looking to a novel causal event but not a novel non-causal event.

In the current study, we adapted the Oakes and Cohen (1990) design in three ways. First, we presented computer-generated clips of inanimate abstract 3D shapes instead of videotaped real objects, allowing us to control the characteristics of the events more carefully. Similar computer-generated stimuli have been used successfully in previous studies with younger infants (e.g., 4-month-olds; Schlottmann, Ray, & Surian, 2012). Second, we implemented a familiarisation procedure in place of habituation. We presented infants with

10

10 trials, each with a maximum duration of 21 seconds. Trials terminated at 21 seconds or earlier if infants looked away for two seconds continuously. Third, we ran only one condition, all infants were familiarised to a non-causal event and then tested on all four event types. In this condition, knowledge of causality is most clearly observed by increased looking to the novel causal but not the novel non-causal event.

### 2.3.1. *Visual stimuli*

Four event sequences were created based on the description in Oakes and Cohen (1990). Events were four seconds in duration and were repeated five times in each presentation. All sequences began with both objects still on screen for one second (Figure 3A). Then object A (e.g., yellow cube) moved towards object B (e.g., green cylinder) for one second. The movement of object B away from object A lasted for one second (Figure 3D). The crucial difference between the events occurred at the mid-point of the sequence. In the direct launching event (causal event), objects A and B touched and object B immediately moved away from A (as if pushed by A; Figure 3B). In the delayed launching event, the objects touched for one second before object B moved away (non-causal event 1; Figure 3B). In the indirect launching event, object A did not touch object B. Object B moved away as object A's movement ended (non-causal event 2; Figure 3C).

Between each sequence, a blank screen was displayed for 200ms. Event clips were created using Processing 2 software (https://processing.org) and portrayed inanimate objects: a yellow cube, green wheel, and a purple hammer in set one, and a blue cylinder, red ball, and green pyramid in set two. The
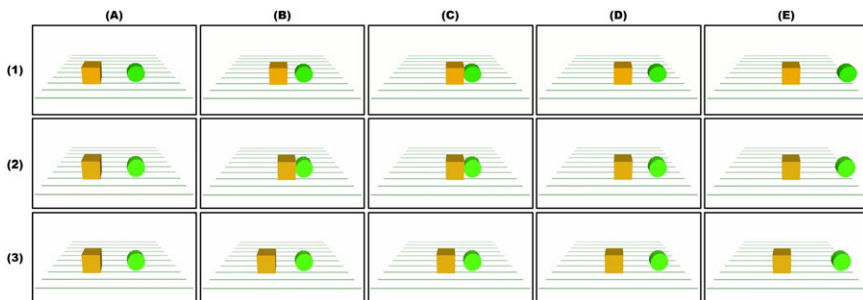


Fig. 3 Screenshots of the caused-motion stimuli showing (1) direct causal launching, (2) delayed non-causal launching, (3) no contact non-causal launching. In the direct launching trials (1), both objects are shown still on screen for 1 second (A). The first object moves towards the second object and stops upon contact (B); the second object then immediately moves away from the first to the other side of the screen (C) and then pauses for 1 second (D). In the delayed non-causal trials (2), there is a 1-second pause after contact (B–C). In the no contact trials (3), the objects never physically connect (C).

11

objects in set one always moved from left to right on the screen and the (non-causal) familiarisation event was indirect launching. The objects in set two always moved from right to left and the (non-causal) familiarisation event was delayed launching. Half of the infants saw the set one objects first and half saw the set two objects first. In individual differences designs it is crucial to minimise item variance to ensure that any variation is due to differences between participants. We implemented a reduced counterbalancing procedure to map onto the counterbalancing in the original design without introducing too much variation into our individual differences design.

### 2.3.2. *Procedure*

Infants were shown 10 familiarisation trials in which a non-causal event was displayed. Looking duration was recorded online; a button was held down by an experimenter watching the infant via a webcam for as long as the infant was attending to the screen. Trials terminated if the infant looked away from the screen for longer than two consecutive seconds. Following familiarisation trials, four test trials began. All infants saw the familiarised non-causal event as the first test trial, followed by the causal event (novel), a new non-causal event (novel), and finally the single object control event (novel). Each test trial had a maximum duration of 21 seconds or ended when the infant looked away from the screen for two seconds continuously.

### 2.4. TASK 2: DISTINGUISHING CHASING FROM RANDOM MOTION

Task 2 was based on the first experiment in Frankenhuis et al. (2013) and was used to determine whether infants show a preference for chasing interactions over random movement sequences involving two coloured circles on the screen. We implemented the multiple cues version of this task, to provide infants with the most complete example of chasing.

### 2.4.1. *Visual stimuli*

The stimuli used in this task were provided by Frankenhuis et al. (2013) and are identical to the stimuli used in their study one. Each video presented two discs, one red and one green, for a fixed duration of 20 seconds. In the chasing video, the discs depicted either the motion of a chaser or of an evader. Both discs had a turning rate of 5% and there was a 5% probability of changing direction with each pixel moved. The chaser disc approached the evader at a constant velocity of .6 pixels per frame in a 'heat-seeking' manner, reducing the distance between the two discs. When this distance was within 80 pixels, the evader would 'flee' by increasing to 4 times its baseline speed (.3 pixels per

12

second) and turn away from the chaser. There were approximately 15 fleeing episodes per 20-second display. Control videos presented the discs moving independently. On occasions where the discs came within 60 pixels of each other, the slower one would change direction but there was no change in velocity and turning rates were lower (.005%). Thus, the videos differed on three dimensions: acceleration, turning rate, and attraction (see Figure 4).

### 2.4.2. *Procedure*

Infants saw 12 trials in total, 6 in each task pairing. On each trial two videos were presented simultaneously side by side in front of the infant; one video depicted random movement (control display) and one chasing movement (target display). Trials played for a fixed duration of 20 seconds. Between trials, a smiley face appeared on screen. Trials were coded offline frame-by-frame (40 frames per second) using Score, part of the Look package (Meints et al., 2008).

### 2.5. TASK 3: IDENTIFYING GOAL-DIRECTED MOVEMENT

The third task we used was an adaptation of Luo (2011) to identify whether infants attribute psychological goals to non-human agents. There were four phases in the task, each ended when the infant looked away for two seconds consecutively or the maximum looking time of 60 seconds was reached. In phase one, infants saw two orientation trials where a single object moved across
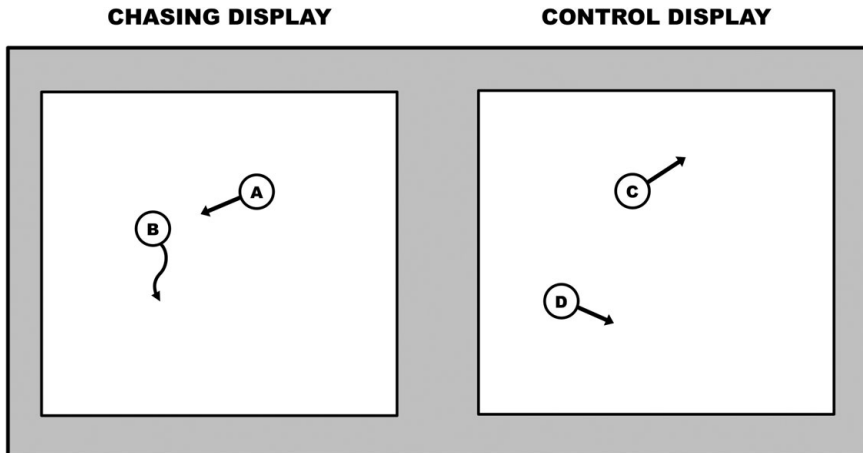


Fig. 4  A diagram of the chasing task. Both event types are presented side by side in a preferential looking task.

13

the screen. In phase two, five familiarisation trials involving three objects, positioned to the right, centre, and left, were presented. The centre object moved towards one of the other objects and paused on contact until the trial ended. Phase three showed a still image where the left- and right-positioned objects were reversed, and the centre stimulus was absent. Phase four consisted of two test trials. In one, the box moved in the opposite direction to familiarisation but towards the same object (old goal event). In the other, the box moved in the same direction but towards the alternative object (new goal event). Luo found that infants looked longer to the new goal event than the old goal event. This was explained as the infants having attributed a psychological goal to the central object's movement towards a particular object.

In our version of this task, we followed the same procedure but made two modifications. First, the familiarisation trials were a fixed duration of eight seconds. Second, computer-generated on-screen stimuli were used instead of real object live-action scenes.

### 2.5.1. *Visual stimuli*

A series of computer-generated scenes involving shapes were created to match, as closely as possible, the scenes described by Luo (2011). The stimulus set 1 presented a blue cone acting as the agent, a red diamond as the target and a yellow box as the distracter (Figure 5F), and stimulus set 2 included a green pyramid as the agent, an orange bowl as the target and a purple cross as the distractor (Figure 5A–E).

Two orientation trials were created. In the first of these, the 'agent' moved from the left of the screen to the centre and back again. In the second, the 'agent' moved from the right of the screen to the centre and back again (Figure 5A). In both cases, the movement cycle lasted six seconds and was shown twice, resulting in a total trial duration of 12 seconds. During familiarisation trials, all three objects were presented with no movement for one second before the 'agent' moved towards the 'target' and away from the 'distractor' (Figure 5B). This movement lasted for three seconds. This final frame, with the 'agent' positioned next to the 'target', remained on screen for a further four seconds. The display event presented a still image of the 'target' and 'distractor' only, but in switched positions (Figure 5C). Finally, two test trials were created. In the old goal event, the 'agent' moved towards the 'target' and remained there for a maximum of 55 seconds (Figure 5E). In the new goal event, the 'agent' moved towards the 'distractor' and remained there for a maximum of 55 seconds (Figure 5D). In both test trials, the shapes were static for one second before the movement began, and the movement lasted for three seconds.
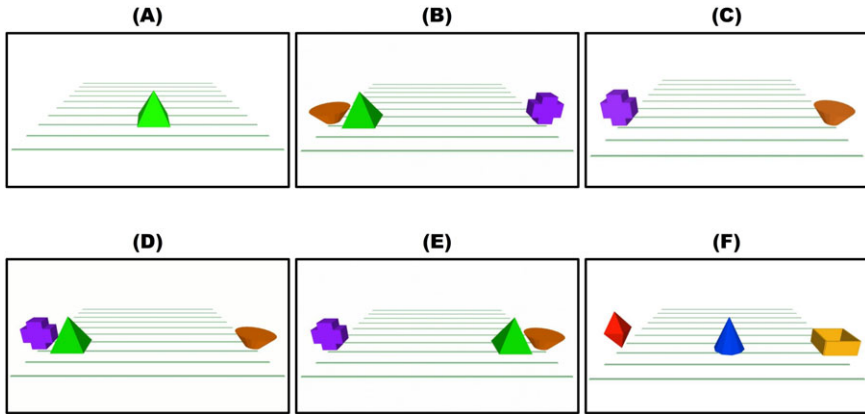
14

Fig. 5 Screenshots of the goals stimuli showing (A) the single object orientation trial, (B) the familiarisation event, (C) the static display with the objects in different positions, (D) the new goal trial, (E) the old goal trial, and (F) the static novel object display.

### 2.5.2. *Procedure*

Looking times were recorded online via a button held down by an experimenter observing the infant via a webcam. Infants were first presented with two orientation trials. Next, infants were shown five familiarisation trials with a fixed duration of eight seconds. Display trials followed the familiarisation and lasted a maximum of 59 seconds, ending when infants looked away for 2 consecutive seconds or the maximum duration was reached. Finally, two test events were presented: in pairing one, the old goal event was presented first, and in pairing two, the new goal event was presented first. Test events lasted a maximum of 59 seconds, or until the infant looked away for 2 seconds consecutively.

### 2.6. VOCABULARY MEASURES

The UK-CDI (Alcock, Meints, & Rowland, 2020) was administered at 8, 9, 11, 12, 15, 16, and 18 months. The UK-CDI is a checklist of the most common gestures and words, which captures UK English-speaking infants' early communicative repertoire and is suitable for use between 9 and 18 months. Caregivers are asked to select 'never', 'sometimes', or 'often' for gesture items and to indicate whether their child 'understands' or 'understands and says' the vocabulary items (total possible score for both receptive and expressive vocabulary = 396). Caregivers were asked to complete the checklist before the visit, either via an online link or on a paper form according to their preference. Those who did not complete it before the visit were given a paper form to complete

15

during the session. The UK-CDI has been standardised for the UK population and has good validity and reliability (see Alcock et al., 2020).

## 2.7. COGNITIVE MEASURES

At 16–17 months the BSID-III (Bayley, 2006) was administered. The BSID-III is made up of five subscales measuring language, cognition, social-emotional, motor, and adaptive behaviour. The cognitive subscale used here comprises 91 items that assess sensorimotor development (e.g., squeezing an object), exploration and manipulation (e.g., finding a hidden object), and pretend play (e.g., demonstrating relational play with toys). The technical manual accompanying this test reports a high degree of reliability (average reliability of .91 for the cognitive scale) and validity (each item within a scale correlated more strongly with that scale than with other scales). Children are awarded a 1 for each correctly performed item.

## 3. Results

### 3.1. GROUP LEVEL ANALYSES

In the first of our planned analyses, we wanted to test whether we replicated the original results in each of the event-understanding tasks. For all three tasks, the dependent measure was the raw looking times in milliseconds on a continuous normal distribution. Linear mixed-effects models were fit to the looking times of each task using the *lme4* 1.1-21 package (Bates, Mächler, Bolker, & Walker, 2015) with the *nloptwrap* optimizer algorithm in *R* version 3.6.1 (R Core Team, 2019). All of the models were initially fitted with the maximal random effects structure (Barr, Levy, Scheepers, & Tily, 2013) but, if the data were unable to support the model, random slopes were removed in a predetermined sequence until convergence was achieved. The highest-order interaction terms were removed first, starting with those that explained the least amount of variance. If the model still did not converge once the interactions were removed, the random slope of *block* was removed before the random slope of *display type*, irrespective of the amount of variance explained by each term, due to the greater theoretical importance of the latter. All of the statistics for the three replication models are presented in Table 1, and the final random effect structure of each model is described in detail below. In addition to the model fit statistics, marginal and conditional $R^2$ are reported as effect sizes. These effect sizes denote the proportion of the variance explained by the model with the inclusion (conditional $R^2$) and exclusion (marginal $R^2$) of the model random effect structure (Johnson, 2014; Nakagawa, Johnson, & Schielzeth, 2017; Nakagawa & Schielzeth, 2013). Bootstrap resampling ($R = 1000$) was used to

TABLE I *The results of the mixed-effect models testing whether the group-level effects were replicated in the three tasks*

| Task | Term | $\beta$ [95% CI] | SE | $\chi^2$ | Pr |
|------|------|-----------------|-----|---------|-----|
| *Causality* | (Intercept) | 6407 [5780, 7008] | 313 | NA | NA |
| | Familiar vs. Non-Causal | −1215 [−2442, 32] | 631 | 0.73 | .536 |
| | Familiar vs. Causal | 1522 [240, 2759] | 643 | 5.61 | .015 |
| | Block 1 vs. Block 2 | −575 [−1019, −122] | 229 | 5.96 | .034 |
| | Familiar/Non-Causal × Block 1/2 | 632 [−643, 1918] | 653 | 0.54 | .463 |
| | Familiar/Causal × Block 1/2 | −437 [−1751, 863] | 667 | 0.46 | .514 |
| *Chasing* | (Intercept) | 5781 [5157, 6423] | 323 | NA | NA |
| | Random vs. Chasing | 1269 [899, 1629] | 186 | 43.06 | < .001 |
| | Block 1 vs. Block 2 | −525 [−716, −335] | 97 | 29.55 | .001 |
| | Random/Chasing × Block 1/2 | −238 [−634, 159] | 202 | 1.56 | .225 |
| *Goals* | (Intercept) | 5720 [5216, 6194] | 249 | NA | NA |
| | Old Goal vs. New Goal | 303 [−584, 1182] | 451 | 0.3 | .563 |
| | Block 1 vs. Block 2 | −199 [−655, 260] | 233 | 0.7 | .401 |
| | Old/New Goal × Block 1/2 | 536 [−336, 1395] | 441 | 1.41 | .250 |

**Causality Model Fit:** BIC = 8022, AICc = 7983, $R_m^2$ = .0291, $R_c^2$ = .1590
**Chasing Model Fit:** BIC = 32924, AICc = 32886, $R_m^2$ = .0380, $R_c^2$ = .1598
**Goals Model Fit:** BIC = 4966, AICc = 4945, $R_m^2$ = .0089, $R_c^2$ = .0747

compute 95% confidence intervals (CIs) for the model estimates and their associated *p*-values (Luke, 2017). The observed effects are illustrated in Figure 6 and the statistics are reported in Table 1.

For the caused-motion task, display type (familiar non-causal, novel causal, novel non-causal) was entered as a centred factor with two contrasts. The first contrast compared looking times for the familiar non-causal trial with the novel non-causal trials, to test whether the infants defamiliarised to a display with different spatio-temporal features but the same non-causal status. The second contrast determined whether the infants defamiliarised to a caused-motion event, comparing looking times in the familiar non-causal trial to the novel causal trial. Additionally, block (1/2) was entered as an effect coded factor, which was crossed with display type to establish whether looking behaviour changed between the first and second presentations. The maximal model supported by the data included *subject* as a random intercept, with the random slope of the familiar/causal contrast. The outcome of this analysis was consistent with the findings of Oakes and Cohen (1990); after familiarisation to a non-causal display, the infants' looking times increased when presented with a novel causal display, but not when presented with a novel non-causal display. Furthermore, while looking times appeared to be higher in block 1 than block 2, there were no interactions between the display contrasts and block,
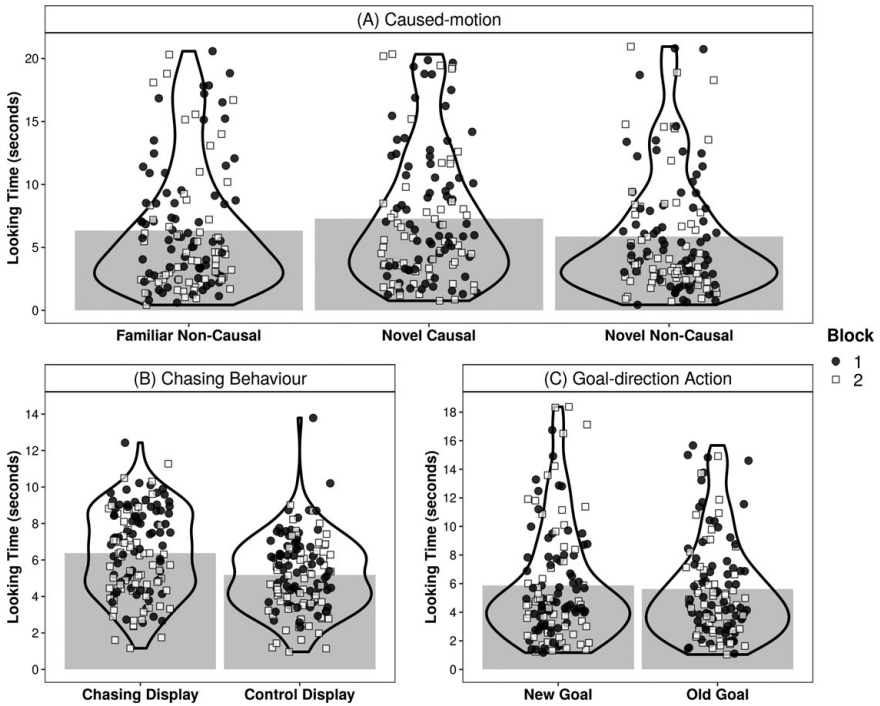
17

Fig. 6 Pirate plots to show the results of the conceptual task replications.

suggesting that the observed group differences in looking times did not vary between the two presentations.

In the chasing task, display type (chasing vs. non-chasing) and block (1/2) were included as fully crossed effect coded fixed factors. Since the chasing task implemented a preferential looking design in which the participants were presented with both display types concurrently over six independent trials, *trial* (1–6) was included as an additional random intercept. The maximal model that converged included both subject and trial as random intercepts but did not include random slopes for these terms. Consistent with the results of Frankenhuis et al. (2013), the model showed that the participants looked for longer at the chasing than the non-chasing display. There was a main effect of block, reflecting a reduction in overall looking time in block 2. However, there was no interaction between block and display type, suggesting that the group preference for the chasing display was consistent.

Finally, in the goal-directed behaviour task, display type (old goal vs. new goal) and block (1/2) were entered into the model as fully crossed effect coded predictors. The random-effects structure supported by the data included

18

subject as a random intercept with the display type contrast as a random slope. The results did not reproduce the findings of Luo (2011); looking duration for trials where the agent approached a new goal was not significantly different from looking duration for trials where it approached the old goal. There was also no difference in overall looking between blocks 1 and 2, nor was there an interaction between block and the display condition, suggesting that looking preferences did not vary between the blocks.

Our version of the goals task made three methodological changes to Luo's (2011) original design, which may have contributed to the difference in results: the participants were 6 to 7 months older than the infants tested by Luo (10 months compared to 2–3 months); we used computer-generated stimuli rather than live object presentation; and we reduced the duration of the familiarisation trials from a maximum of 60 seconds to a fixed 8-second period. To rule out this third change as an explanation for our failure to reproduce Luo's result, we ran a closely matched follow-up study with a separate sample of 26 infants (15 girls) aged 9 to 12 months (*Median* = 321 days, *Range* = 274–369 days). This version used the same stimuli as our first reproduction attempt but implemented an infant-controlled familiarisation phase closely aligned to Luo's procedure. In this version of the task, the familiarisation trials could last up to 60 seconds. Trials were terminated either when the child looked away for 2 seconds continuously or the maximum duration was reached. The results of this second experiment did not reproduce Luo's findings either, but were consistent with the results of our first task (see online materials at <https://osf.io/mjv73/>). In Luo's work, large effect sizes were reported for the differences in looking times between the new goal and old goal trials (exp.1: *Cohen's d* = 0.8; exp.2: *Cohen's d* = 0.7). Using bootstrap simulations ($R$ = 10000), confidence intervals were computed for the Cohen's d observed in our two reproduction attempts. The observed effect sizes were considerably smaller and the CIs did not capture the effects reported in Luo's study (attempt 1: *Cohen's d* = –0.07 [–0.31, 0.18]; attempt 2: *Cohen's d* = 0.07 [–0.34, 0.49]). In sum, it appears that the results of the caused-motion and chasing tasks reproduced the group effects reported by the original authors (Frankenhuis et al., 2013; Oakes & Cohen, 1990) while the goal-directed behaviour task did not (Luo, 2011).

### 3.2. INDIVIDUAL DIFFERENCES AND RELATIONSHIPS WITH VOCABULARY DEVELOPMENT

Our second set of planned analyses assessed whether there are meaningful individual differences in 9- to 10-month-olds' ability to interpret dynamic events and whether these differences influence vocabulary development. The difference in looking times to the target and control displays in each task was

19

used as a measure of infants' understanding. For the caused-motion task, this was total looking time in the novel causal trial minus total looking time in the novel non-causal trial. For the chasing task, it was the looking time to the chasing display minus looking time to the random movement display. For the goal-directed behaviour task, it was the total looking time during the new goal trial minus total looking time in the old goal trial. We performed a series of Pearson's correlations that were bootstrapped ($R = 10000$) to obtain 95% confidence intervals and accurate $p$-values. We can infer that there is a meaningful direction-specific relationship between two measures when they produce a high Pearson's $r$ coefficient, 95% CIs that do not cross zero, and a low $p$-value.

We first examined whether the infants' performance in block 1 correlated with their performance in block 2 within each task. These correlations found no evidence of consistency; all three tasks yielded 95% CIs that crossed zero and had high $p$-values (all $p$s > .217, see Table 2). This suggests that, at the individual level, the infants responded differently in each presentation of the three tasks. This lack of consistency may be due to fatigue effects since overall looking times were lower in the second block for both the caused-motion and chasing tasks, or to carry-over effects from having seen similar displays, albeit with different objects, in block 1. It may also be that the looking-time difference scores were not representative of individual differences in understanding. It is not possible to disentangle these explanations on the basis of the current data but, considering these results, all subsequent analyses were performed using the data from each block separately.

TABLE 2 *Looking-time correlations for each task across blocks 1 and 2*

|  | $x$ | $y$ | Pearson's [95% CI] | Df | Pr |
|---|---|---|---|---|---|
| Within Task | Causality 1 | Causality 2 | .04 [−.18, .28] | 56 | .526 |
|  | Chasing 1 | Chasing 2 | −.19 [−.39, .03] | 69 | .217 |
|  | Goals 1 | Goals 2 | −.11 [−.39, .17] | 49 | .406 |
| Between Tasks | Causality 1 | Chasing 1 | −.26 [−.45, −.02] | 66 | .116 |
|  | Causality 1 | Chasing 2 | .13 [−.12, .40] | 65 | .350 |
|  | Causality 1 | Goals 1 | −.20 [−.44, .05] | 61 | .226 |
|  | Causality 1 | Goals 2 | −.01 [−.21, .21] | 50 | .587 |
|  | Chasing 1 | Goals 1 | −.02 [−.24, .17] | 62 | .565 |
|  | Chasing 1 | Goals 2 | .10 [−.32, .40] | 52 | .350 |
|  | Causality 2 | Chasing 1 | −.02 [−.35, .18] | 57 | .509 |
|  | Causality 2 | Chasing 2 | −.21 [−.39, .01] | 57 | .185 |
|  | Causality 2 | Goals 1 | −.10 [−.41, .15] | 55 | .417 |
|  | Causality 2 | Goals 2 | −.06 [−.46, .28] | 48 | .379 |
|  | Chasing 2 | Goals 1 | −.09 [−.32, .14] | 62 | .428 |
|  | Chasing 2 | Goals 2 | .05 [−.17, .30] | 52 | .529 |

Our next analysis tested whether the infants' looking behaviour was driven by a common understanding of dynamic events shared between the tasks. Following the same analytical procedure as above, we tested whether the infants showed a consistent preference direction and magnitude for the target/control display across the three tasks. However, these analyses showed no reliable relationships in looking behaviour in either block (see Table 2). The infants who, for example, strongly discriminated causal from non-causal motion events were not the same individuals who demonstrated strong preferences in the chasing or goal-directed behaviour tasks.

Finally, we tested whether the separate abilities measured by the tasks were related to vocabulary development. The infants' task difference scores were combined with three separate measures of vocabulary obtained from the UK-CDI questionnaire at six different age points (9, 11, 12, 15, 16, and 18 months), total comprehension vocabulary, total productive vocabulary, and action word comprehension. Correlations were calculated for both the block 1 and block 2 difference scores following the same procedure as the previous analyses. The purpose of these correlations was not to directly test our hypotheses, but to identify whether a more rigorous confirmatory growth curve analysis (GCA) was merited. This would provide a method of estimating change in vocabulary size over time as a function of individual differences in the looking-time difference scores (Mirman, 2014; Mirman, Dixon, & Magnuson, 2008). Specifically, the correlations served as preliminary checks for potential relationships in the data to identify the parameters to include in a full GCA model (e.g., the caused-motion but not the goal-directed behaviour scores).

However, these correlation analyses found no consistent relationships between the task and vocabulary measures (see Tables 3 and 4). The strongest relationships were between the chasing task and the total comprehension and action word comprehension scores, but the effects occurred in opposite directions in each block and none of the $p$-values reached the traditional significance threshold (cf. McShane, Gal, Gelman, Robert, & Tackett, 2019). One should also consider that a total of 162 correlations were performed, so approximately eight correlations with a traditionally significant $p$-value would be expected to occur by chance under the null hypothesis. The main implication of these analyses is that there was no strong evidence of a relationship between vocabulary development and looking-time preferences in the three event perception tasks.

### 3.3. EXPLORATORY ANALYSES AND SIMULATIONS

The present results suggest that visual event understanding at 10 months does not predict concurrent vocabulary size and ongoing vocabulary development.

21

TABLE 3 Pearson's *R* [95% CIs] to show the correlations between task performance in block 1 and the target UK-CDI vocabulary scores

| Task | Months | All Receptive | All Expressive | Action Receptive |
|---|---|---|---|---|
| *Causality* | 9 | −.15 [−.32, .03] | .04 [−.18, .24] | −.08 [−.28, .13] |
| | 11 | −.06 [−.29, .13] | .16 [−.04, .34] | −.08 [−.30, .11] |
| | 12 | −.18 [−.48, .03] | .06 [−.11, .25] | −.19 [−.52, .03] |
| | 15 | −.06 [−.39, .17] | .14 [−.02, .27] | −.08 [−.36, .14] |
| | 16 | −.05 [−.34, .18] | .07 [−.19, .26] | −.05 [−.30, .17] |
| | 18 | −.01 [−.26, .23] | .02 [−.21, .23] | .01 [−.24, .22] |
| *Chasing* | 9 | .19 [.02, .36] | −.05 [−.27, .21] | .22 [−.03, .35] |
| | 11 | .23 [.04, .39] | −.11 [−.33, .11] | .22 [.00, .35] |
| | 12 | .26 [.05, .45] | .07 [−.07, .21] | .26 [.09, .45] |
| | 15 | .07 [−.19, .33] | .02 [−.21, .19] | .07 [−.15, .31] |
| | 16 | .01 [−.26, .31] | −.04 [−.27, .16] | .05 [−.30, .31] |
| | 18 | −.05 [−.33, .27] | −.14 [−.35, .09] | −.10 [−.33, .14] |
| *Goals* | 9 | −.06 [−.34, .19] | −.09 [−.27, .11] | −.11 [−.32, .14] |
| | 11 | −.00 [−.19, .22] | −.02 [−.20, .20] | −.01 [−.19, .22] |
| | 12 | −.06 [−.29, .17] | −.03 [−.32, .24] | −.08 [−.31, .18] |
| | 15 | .01 [−.22, .24] | .02 [−.17, .20] | −.05 [−.28, .18] |
| | 16 | .04 [−.19, .27] | −.02 [−.20, .15] | .01 [−.22, .24] |
| | 18 | −.06 [−.26, .13] | .02 [−.18, .22] | −.02 [−.23, .17] |

* *p* < .05, *** *p* < .001

TABLE 4 Pearson's *R* [95% CIs] to show the correlations between task performance in block 2 and the target UK-CDI vocabulary scores

| Task | Months | All Receptive | All Expressive | Action Receptive |
|---|---|---|---|---|
| *Causality* | 9 | .04 [−.21, .26] | .24 [−.03, .44] | .10 [−.14, .34] |
| | 11 | .21 [−.05, .39] | .14 [−.12, .32] | .20 [−.06, .36] |
| | 12 | .19 [−.03, .39] | .12 [−.08, .27] | .22 [−.01, .41] |
| | 15 | .07 [−.24, .32] | .06 [−.15, .24] | .02 [−.32, .28] |
| | 16 | −.01 [−.27, .25] | −.03 [−.26, .15] | .01 [−.29, .27] |
| | 18 | .10 [−.14, .33] | .01 [−.15, .20] | .17 [−.11, .40] |
| *Chasing* | 9 | −.24 [−.53, .08] | −.11 [−.45, .19] | −.34 [−.57, −.10] |
| | 11 | −.29 [−.56, .02] | −.21 [−.53, .06] | −.29 [−.54, .01] |
| | 12 | −.37 [−.61, −.05] | −.18 [−.49, .09] | −.32 [−.57, −.01] |
| | 15 | −.32 [−.56, −.04] | −.11 [−.41, .07] | −.32 [−.53, −.07] |
| | 16 | −.25 [−.48, .02] | −.17 [−.42, .05] | −.23 [−.45, .04] |
| | 18 | −.21 [−.44, .06] | −.18 [−.42, .07] | −.21 [−.45, .04] |
| *Goals* | 9 | −.01 [−.20, .15] | −.03 [−.25, .17] | −.02 [−.21, .15] |
| | 11 | −.04 [−.29, .13] | −.05 [−.28, .16] | −.03 [−.30, .15] |
| | 12 | .00 [−.20, .18] | .09 [−.13, .41] | .00 [−.20, .23] |
| | 15 | −.14 [−.37, .09] | .03 [−.27, .43] | −.10 [−.32, .14] |
| | 16 | −.04 [−.29, .24] | .06 [−.27, .54] | −.03 [−.29, .22] |
| | 18 | .05 [−.27, .34] | .10 [−.17, .46] | −.02 [−.32, .28] |

* *p* < .05, *** *p* < .001

Additional exploratory analyses were performed on the block 1 data for the caused-motion and chasing tasks. These two datasets were selected since both tasks replicated the original group-level effects, while also providing examples from familiarisation and preferential-looking designs. Although the results of our planned analyses found no evidence of a relationship between vocabulary size and the understanding of dynamic events, they made two critical assumptions about how such knowledge would be expressed. First, it was assumed that this knowledge exists on an interval scale that is directly expressed through the infants' looking times. For example, an infant with an 800ms preference for the chasing display is considered twice as knowledgeable as an infant with a 400ms preference. The second assumption is that the direction of the preference is meaningful; that is, an infant with a 200ms preference for the novel causal display is treated as having more knowledge about caused-motion events than an infant with a 200ms preference for the novel non-causal display (i.e., a difference score of −200ms). We tested the validity of these assumptions in a series of exploratory analyses and simulations.

The first simulation considered whether an individual infant's dynamic event understanding should be expressed categorically rather than continuously. The looking-time difference scores calculated for the planned analyses were used to classify each infant as having either a target display preference (i.e., longer looking at novel causal or chasing display), a control display preference (i.e., longer looking at the novel non-causal or random motion display), or no preference. These categories represented levels on a *response type* factor, coded with Helmert contrasts. The first contrast compared the infants with a target preference to those with a control preference, to explore whether categorical differences in response direction may predict vocabulary development. The second contrast pooled the infants showing either a target or control preference and compared them to the group showing no preference. This comparison considers whether the infants who make a distinction between the two display types, regardless of the stimulus they prefer, have a larger vocabulary size than those who looked at both displays equally.

When converting a continuous variable into a categorical factor, it is necessary to determine the boundaries of each category. To avoid making assumptions about where this threshold occurs, we simulated a range of boundaries by iteratively raising the no-preference limit in looking-time increments representing 1% of the standard deviation of each sample (caused-motion increments = 63ms; chasing increments = 24ms). Both the caused-motion and chasing simulations were initialised with a no-preference upper boundary at SD × 0.05 (i.e., a *very small* effect size; Sawilowsky, 2009) and terminated at SD × 0.5 (i.e., a *medium* effect size; Cohen, 1992). With each iteration, separate bootstrapped linear regression models ($r = 1000$) were fit to the total comprehension scores for each of the six UK-CDI age-points (9, 11, 12, 15, 16, and

23

18 months), with the *response type* factor as a predictor. These simulations found no evidence that non-linear looking preferences were related to comprehension vocabulary; the *t*-ratios for each contrast were small and inconsistent across the different ages in every iteration (see online materials at <https://osf.io/mjv73/>).

The second simulation considered the possibility that looking-time differences are only partially related to vocabulary size; that there may be a difference threshold where the relationship diminishes. To explore whether such a threshold exists, a second simulation iteratively examined the relationship between looking times and total comprehension vocabulary in participants who had difference scores below a set threshold. In a similar procedure to the first simulation, the inclusion limit was incrementally increased by 1% of the standard deviation for the entire sample, thus more data were included with each iteration. For both tasks, the simulations were initialised with an inclusion threshold of SD × 0.05 and terminated at SD × 2.5. For every iteration, bootstrapped linear regression models ($r = 1000$) were fit to the total comprehension vocabulary for each of the six assessed ages. Unlike the planned analyses, preference direction was entered as a separate predictor from the continuous looking-time scores. Specifically, the absolute value of the infants' looking times was taken as a measure of their response magnitude, with the direction of this score representing their stimulus preference as an effect coded factor. Consistent with the results of the first simulation, the analyses produced no evidence of a threshold; small and variable *t*-ratios were observed throughout the analysis (see online materials at <https://osf.io/mjv73/>).

A third simulation assessed whether there was a relationship between dynamic event understanding and vocabulary development that includes both linear and non-linear components. The previous simulation examined whether there is a linear relationship between event understanding and vocabulary that disappears beyond a certain threshold. This third simulation considered whether this threshold marks a transition in the relationship from linear to categorical. Following a similar procedure to the previous simulations, the switch point was incrementally increased by 1% of the standard deviation of the sample, starting at a threshold of SD × 0.2 and terminating at a limit equal to SD × 1. This threshold was used to split the data into linear and non-linear components. The proportion of the dataset being treated as linear increased as the simulation progressed. When a difference score exceeded the limit, its value was replaced with the mean looking time of the data above the threshold. Thus, all looking behaviour above the threshold was treated as statistically equivalent. No transformations were applied to difference scores that fell below the threshold. For each iteration, the data were fit with the same bootstrapped linear regression models as the second simulation ($r = 1000$), with separate predictors for difference scores and preference type. Consistent with the

24

previous simulations, these analyses found no evidence to support a non-linear switching point in the relationship between looking times and vocabulary (see online materials at <https://osf.io/mjv73/>).

We found that 10-month-old infants can distinguish between displays showing caused-motion or non-causal events, and between chasing interactions and random movement. However, their precise looking behaviour in these tasks appears to have no meaningful relationships with vocabulary development (measured by the UK-CDI) or other tasks measuring different event perception abilities. Given these results, it is possible that looking behaviour in these familiarisation and preferential looking tasks does not directly correspond to the infants' *depth* of understanding of the stimulus and is not suitable for measuring individual differences. In our fourth exploratory analysis, we considered whether the Pearson's *r* coefficients observed in our planned analyses (see Tables 3 and 4) could be replicated in a simulation that treated looking times as meaningful effects at the group level, but randomly distributed at the individual level. Consistent with the previous analyses, only the block 1 data from the caused-motion and chasing tasks were used. For each iteration, the data were randomly reshuffled at the trial level so that each participant's task difference scores no longer corresponded to their vocabulary scores. By reshuffling the data, we were able to retain the group effects observed in the replication analyses, while concurrently assessing whether the individual looking times that produced these effects have a meaningful relationship with vocabulary size. For each of the six UK-CDI age-points (9, 11, 12, 15, 16, and 18 months), the caused-motion and chasing difference scores were reshuffled 10,000 times and a Pearson's correlation was computed with the mismatched total comprehension scores. This produced a distribution of Pearson's *r* coefficients, which was used to calculate the probability of the correlations observed in our planned analyses, or larger, under the hypothesis that looking times were truly random. The outcome of these simulations (see Figures 7 and 8) suggests that many of these correlations have a high probability of occurrence under the computed distribution, especially those for the later assessment ages. Some of the early correlations with the chasing task showed a low probability but remained possible under the simulated distribution. Therefore, many of the relationships between looking-time scores and vocabulary have a medium to high probability of occurrence if the task measures do not reliably capture any meaningful individual differences in the understanding of dynamic events.

The fifth analysis tested whether the problem might lie in the looking-time measures themselves. Since the looking-time differences used in the analyses produced no reliable relationships, it is possible that these scores were unsuitable as measures of individual differences in event understanding. Thus, the same analytical procedure used in the UK-CDI vocabulary analyses above was
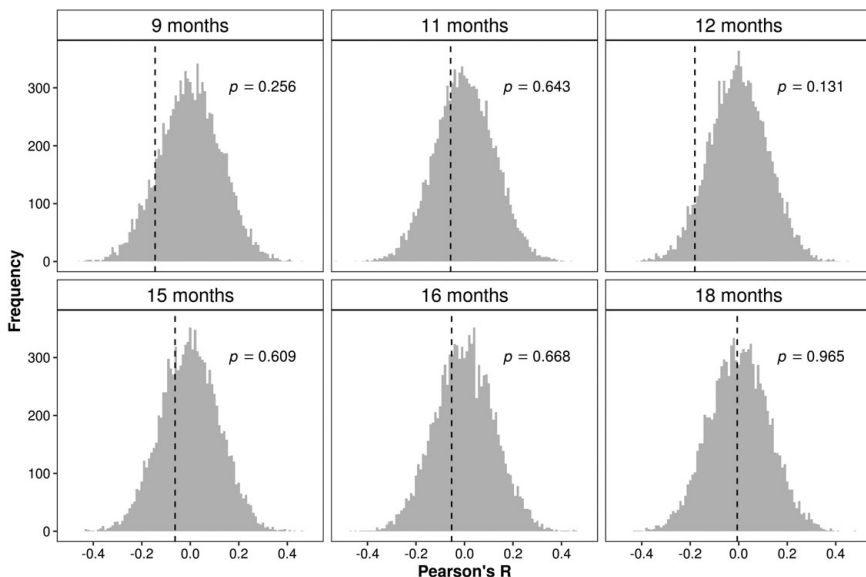
25

Fig. 7 Histograms of the correlation coefficients computed in the causality random reshuffle simulation. The dashed lines represent the observed correlation coefficients with their *p*-values based on the generated distribution.

used to examine whether there was a relationship between the looking-time difference scores extracted from the three visual tasks and the BSID-III cognitive development subscale administered when the infants were aged 16–17 months, on the basis that an individual's event understanding might reasonably be expected to correlate with global cognitive ability. However, no reliable relationships were observed between the task looking times and cognitive development scores (see Table 5).

Finally, a sixth analysis examined whether an alternative measure derived from these tasks would be more reliable. Since many studies have reported that the rate of habituation is stable within individuals over time (e.g., Bornstein & Benasich, 1986; Colombo, Mitchell, Coldren, & Freeseman, 1991; McCall & Carriger, 1993), the infants' total looking times during the familiarisation stages of the caused-motion and goal-directed behaviour tasks were calculated and correlated across blocks 1 and 2 within each task. The scores for each block were then individually correlated with the three measures of vocabulary size (total comprehension, total production, and action words) at all six assessment age-points. There was no evidence of individual consistency within each task, nor were there any notable relationships with vocabulary (see online materials at <https://osf.io/mjv73/>), suggesting that the absence of individual
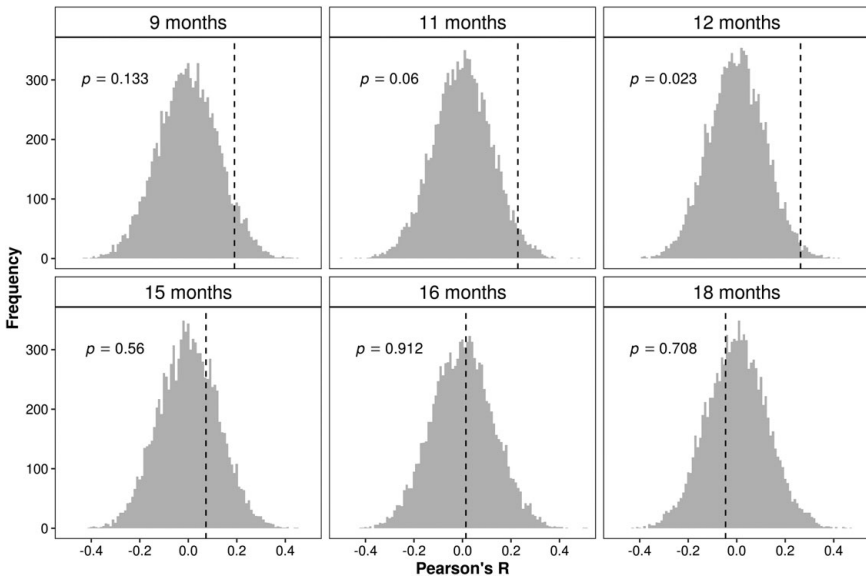
26

Fig. 8 Histograms of the correlation coefficients computed in the chasing random reshuffle simulation, with dashed lines to show the observed correlation coefficients and their *p*-values on the simulated distribution.

TABLE 5 *Correlations between task performance and Bayley cognitive development scores at 16 months*

| Block | Task | Pearson's [95% CI] | Df | Pr |
|---|---|---|---|---|
| 1 | *Causality* | .16 [−.14, .42] | 62 | .293 |
| | *Chasing* | −.13 [−.44, .15] | 63 | .337 |
| | *Goals* | −.02 [−.25, .29] | 60 | .473 |
| 2 | *Causality* | −.19 [−.51, .08] | 52 | .285 |
| | *Chasing* | .13 [−.20, .39] | 61 | .344 |
| | *Goals* | .02 [−.31, .37] | 46 | .434 |

differences effects in the present work cannot be attributed to the specific measures entered into these analyses.

## 4. General discussion

Most theories of language acquisition assume that visual–spatial perceptual processing is important for constructing shared language meaning (Goldin-Meadow & Mylander, 1998; Landau & Jackendoff, 1993; Mandler, 1992; Tomasello & Vaish, 2013). Without this ability, there would be nothing to

27

ensure that the meaning of "chase" was similar across individuals. The goal of the present research was to examine this assumption using a longitudinal individual differences design. We tested whether there were meaningful individual differences in the ability of 9- to 10-month-old infants to interpret three dynamic events, and whether these differences influenced vocabulary acquisition. Ours is the first study to explicitly test the relationship between conceptual development and language acquisition at the age when language is starting to be learned by children (i.e., just before their first birthday), and in multiple conceptual understanding tasks.

Analysis one was a conceptual replication of three tasks that assessed dynamic event understanding. Our 9- to 10-month-olds could distinguish between causal and non-causal movement, consistent with the findings of Oakes and Cohen (1990) and others (Leslie & Keeble, 1987). They also demonstrated a preference for chasing events over a concurrently presented display showing random movement, consistent with Frankenhuis et al. (2013). In contrast, across two experiments, we found no evidence that 9- to 10-month-olds infer object preferences in the movement of abstract shapes (Luo, 2011). There are two potential explanations for this discrepancy. One is the age difference; Luo tested 2- and 3-month-olds, so our 9- to 10-month-olds may have found the task too easy and extracted the relevant information from the scenes very quickly. This is likely to mean that the looking times do not reflect their knowledge. The second explanation concerns the stimuli. We used on-screen animated objects, which showed movement with a smooth consistent velocity. Luo used physical objects in a wooden box display, with the central agent controlled by a human experimenter. It is possible that human experimenters unconsciously varied the speed of motion of the objects, which could create a greater impression of animacy, helping the infants to interpret the behaviour as goal-directed. Further work is needed to address these issues.

Our second and third set of analyses tested whether there were reliable and correlated individual differences in looking behaviour across the three tasks, and whether these differences were related to vocabulary size. We observed no consistency in individual performance on the same task at the two time-points, no relationships in looking behaviour between the three different tasks, and nor was there any evidence of a link between task performance and vocabulary or cognitive ability. Although the failure to observe consistency across the two time-points (blocks) might be explained by fatigue (due to the duration of the session) or transfer effects (since the infants watched similar displays, albeit with different objects, in both blocks), this cannot adequately explain the failure to find a relationship across tasks, or with vocabulary. The looking times in the second block of each task were shorter; however, the overall pattern of results was consistent. The group effects were present in both blocks and the relationships with language and between tasks were absent. Our exploratory

28

analyses confirmed the absence of such meaningful relationships in our data. We used statistical simulations to consider different explanations about why we might have failed to observe relationships between vocabulary development and our looking-time measures: first, that knowledge of causal relationships is categorical; second, that the magnitude of the difference and not its direction is reflecting knowledge; third, that there is a threshold beyond which knowledge is no longer captured by looking-time differences; and finally, that the relationship is non-linear or has non-linear components. We found no evidence of any relationships between these tasks and language development within any of our simulations. Indeed, our fourth simulation revealed that our findings were highly probable on a distribution of Pearson's $r$ coefficients computed by randomly shuffling the infants' looking times and correlating them with vocabulary. In our fifth exploratory analysis, we determined if the individual looking times derived from the tasks relate to global cognitive ability. If such a relationship with cognitive development had been observed, this would have provided clear evidence that the looking-time measures we were using had indeed captured variation in infants' knowledge of direct causality, agency, and intention. However, since we did not observe any relationships, using looking times from these tasks as measures of knowledge may be inappropriate. Thus, in our sixth and final exploratory analysis, we tested whether another measure would provide more reliable individual level scores: infants' total looking time during familiarisation. There was no change in the pattern of results.

Considering the cumulative findings of these analyses, there are three possible explanations for the present results. The first is that looking-time differences in these tasks are meaningful, but each is capturing a different ability, so we fail to see relationships between these tasks. The second, not mutually exclusive, explanation is that looking times are meaningful but these abilities have no effect on children's vocabulary learning. On this account, although children do need to be able to interpret events to learn the associated words, the necessary knowledge has already been acquired by the earliest age we measured their vocabulary (i.e., 9–10 months). However, a third explanation is that looking-time differences are not a reliable measure of an individual infant's understanding. We consider this final explanation to be the most plausible, in the light of the results of our a priori and exploratory analyses.

This conclusion raises broader questions about how to use looking times as a measure of individual variation (see Hedge, Powell, & Sumner, 2017). One question concerns experimental design. To find a group effect, we need to minimise between-subject variance, but to capture individual differences, we need high between-subject variance (Hedge et al., 2017). Thus, the very methodological decisions that lead to success in group designs may reduce our chances of identifying meaningful individual differences. Furthermore, it should be noted that in the caused-motion and goal-directed behaviour tasks,

29

the looking-time differences were derived from a single trial of each display (e.g., novel causal vs. novel non-causal). Prior work that has found relationships between looking-time measures and other aspects of development (e.g., IQ; McCall & Carriger, 1993) have used measures over multiple presentations of the same stimuli. Thus, those looking times may be less influenced by attentional and environmental factors than the ones used in these analyses. That said, the chasing task and our exploratory analyses using rate of familiarisation both calculated individual differences aggregated over multiple trials. Still, neither of these produced evidence of relationships within the tasks, between the tasks, or with vocabulary.

Another question concerns the age of testing. For infants to demonstrate success in a group design, most of the group need to show an effect in the predicted direction. However, in an individual differences design, we are looking for the highest variance possible, which may be when group performance is at chance. Thus, at the age at which we find group effects, individual variance may no longer be large enough to measure meaningful differences in performance. In our case, if we had tested infants younger than 9–10 months of age, and managed to capture the transitionary period of the development of causal understanding, individual differences might have emerged.

A third question concerns the fact that we know very little about the underlying drivers of looking behaviour in these tasks, which means it is difficult to make predictions on an individual level. Looking-time difference measures are based on the premise that children with a more mature understanding of dynamic events will show bigger looking differences (i.e., will look for substantially longer at one event than the other) than those with immature or emerging knowledge. Yet, an infant with a more mature understanding may, in fact, extract the relevant information in the events quickly, and then spend the rest of the time looking at both displays. Alternatively, an infant with an immature or emerging concept of causality may take longer to extract relevant information, and thus may explore one scene more than the other. In other words, the underlying strategies that drive looking-time behaviour may not yield the expected results. Much more work is needed to understand the relationship between looking-time behaviour and the knowledge and processes that drive this behaviour in infants.

## 5. Conclusion

In summary, we successfully replicated the results of two of the three tasks: Oakes and Cohen's caused-motion task and Frankenhuis' chasing task, but not Luo's goal-directed behaviour task. Planned analyses revealed no relationships between these tasks or with vocabulary size. Exploratory analyses also found no relationship with a measure of cognitive development. Simulations explored

four explanations and converged on a consensus that there were no relationships with language within our data. These findings raise important questions about how to convert experimental group designs into individual differences designs, and how to interpret infant looking-time behaviour. We propose that differences in infants' ability to interpret dynamic events at 9–10 months of age are not causally implicated in explaining individual variation in vocabulary acquisition, at least in typically developing children.

REFERENCES

Alcock, K. J., Meints, K. & Rowland, C. F. (2020). *The UK Communicative Development Inventories: words and gestures*. Guilford: J&R Press Ltd.

Astington, J. W. (1986). Children's comprehension of expressions of intention. *British Journal of Developmental Psychology* **4**(1), 43–49.

Baillargeon, R., Scott, R. M. & Bian, L. (2016). Psychological reasoning in infancy. *Annual Review of Psychology* **67**(1), 159–186.

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

Barrett, H. C., Todd, P. M., Miller, G. F. & Blythe, P. W. (2005). Accurate judgments of intention from motion cues alone: a cross-cultural study. *Evolution and Human Behavior* **26**(4), 313–331.

Bates, D., Mächler, M., Bolker, B. & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* **67**(1), 1–48.

Bayley, N. (2006). *Bayley Scales of Infant and Toddler Development*. San Antonio, TX: PsychCorp, Pearson.

Bornstein, M. H. & Benasich, A. A. (1986). Infant habituation: assessments of individual differences and short-term reliability at five months. *Child Development* **57**(1), 87–99.

Bornstein, M. H. & Sigman, M. D. (1986). Continuity in mental development from infancy. *Child Development* **57**(2), 251–274.

Chang, F., Choi, Y. & Ko, Y. (2015). Why loose rings can be tight: the role of learned object knowledge in the development of Korean spatial fit terms. *Cognition* **136**, 196–203.

Choi, S. & Bowerman, M. (1991). Learning to express motion events in English and Korean: the influence of language-specific lexicalization patterns. *Cognition* **41**(1/3), 83–121.

Cohen, J. (1992). A power primer. *Psychological Bulletin* **112**(1), 155–159.

Cohen, L. B. & Amsel, G. (1998). Precursors to infants' perception of the causality of a simple event. *Infant Behavior and Development* **21**(4), 713–731.

Cohen, L. B. & Oakes, L. M. (1993). How infants perceive a simple causal event. *Developmental Psychology* **29**(3), 421–433.

Colombo, J., Mitchell, D. W., Coldren, J. T. & Freeseman, L. J. (1991). Individual differences in infant visual attention: Are short lookers faster processors or feature processors? *Child Development* **62**(6), 1247–1257.

Csibra, G. (2008). Goal attribution to inanimate agents by 6.5-month-old infants. *Cognition* **107**(2), 705–717.

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**(3), 237–267.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J. … Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development* **59**(5), i–185.

31

Fernald, A., Perfors, A. & Marchman, V. A. (2006). Picking up speed in understanding: speech processing efficiency and vocabulary growth across the 2nd year. *Developmental Psychology* **42**(1), 98–116.

Frank, M. C., Braginsky, M., Yurovsky, D. & Marchman, V. A. (2017). Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language* **44**(3), 677–694.

Frankenhuis, W. E., House, B., Barrett, H. C. & Johnson, S. P. (2013). Infants' perception of chasing. *Cognition* **126**(2), 224–233.

Galazka, M. & Nyström, P. (2016). Infants' preference for individual agents within chasing interactions. *Journal of Experimental Child Psychology* **147**, 53–70.

Gampe, A., Keitel, A. & Daum, M. M. (2015). Intra-individual variability and continuity of action and perception measures in infants. *Frontiers in Psychology* **6**, e2015.00327.

Gao, T., Newman, G. E. & Scholl, B. J. (2009). The psychophysics of chasing: a case study in the perception of animacy. *Cognitive Psychology* **59**(2), 154–179.

Gao, T. & Scholl, B. J. (2011). Chasing vs. stalking: interrupting the perception of animacy. *Journal of Experimental Psychology: Human Perception and Performance* **37**(3), 669–684.

Gathercole, S. E. & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: a longitudinal study. *Journal of Memory and Language* **28**(2), 200–213.

Gergely, G., Nádasdy, Z., Csibra, G. & Bíró, S. (1995). Taking the intentional stance at 12 months of age. *Cognition* **56**(2), 165–193.

Goldin-Meadow, S. & Mylander, C. (1998). Spontaneous sign systems created by deaf children in two cultures. *Nature* **391**(6664), 279–281.

Hedge, C., Powell, G. & Sumner, P. (2017). The reliability paradox: why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods* **50**(3), 1166–1186.

Hespos, S. J. & Spelke, E. S. (2004). Conceptual precursors to language. *Nature* **430**(6998), 453–456.

Hoff, E. (2006). How social contexts support and shape language development. *Developmental Review* **26**(1), 55–88.

Hume, D. (1748). *An enquiry concerning human understanding* (T. L. Beauchamp, ed.). Oxford: Clarendon Press

Johnson, P. C. D. (2014). Extension of Nakagawa & Schielzeth's $R^2_{\text{GLMM}}$ to random slopes models. *Methods in Ecology and Evolution* **5**(9), 944–946.

Kavšek, M. (2004). Predicting later IQ from infant visual habituation and dishabituation: a meta-analysis. *Journal of Applied Developmental Psychology* **25**(3), 369–393.

Kidd, E., Junge, C., Spokes, T., Morrison, L. & Cutler, A. (2018). Individual differences in infant speech segmentation: achieving the lexical shift. *Infancy* **23**(6), 770–794.

Landau, B. & Jackendoff, R. (1993). 'What' and 'where' in spatial language and spatial cognition. *Behavioral and Brain Sciences* **16**(2), 217–238.

Langacker, R. W. (1987). *Foundations of cognitive grammar: theoretical prerequisites*. Stanford, CA: Stanford University Press.

Lany, J. (2018). Lexical-processing efficiency leverages novel word learning in infants and toddlers. *Developmental Science* **21**(3), e12569.

Leslie, A. M. (1982). The perception of causality in infants. *Perception* **11**(2), 173–186.

Leslie, A. M. (1984). Spatiotemporal continuity and the perception of causality in infants. *Perception* **13**(3), 287–305.

Leslie, A. M. & Keeble, S. (1987). Do six-month-old infants perceive causality? *Cognition* **25**(3), 265–288.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behavior Research Methods* **49**(4), 1494–1502.

Luo, Y. (2011). Three-month-old infants attribute goals to a non-human agent: three-month-olds attribute goals to a non-human agent. *Developmental Science* **14**(2), 453–460.

Luo, Y. & Baillargeon, R. (2005). Can a self-propelled box have a goal? Psychological reasoning in 5-month-old infants. *Psychological Science* **16**(8), 601–608.

Mandler, J. M. (1992). How to build a baby: II. *Conceptual primitives*. Psychological Review **99**(4), 587–604.

32

Mani, N. & Huettig, F. (2012). Prediction during language processing is a piece of cake – but only for skilled producers. *Journal of Experimental Psychology: Human Perception and Performance* **38**(4), 843–847.

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. San Francisco, CA: W. H. Freeman and Company.

McCall, R. B. & Carriger, M. S. (1993). A meta-analysis of infant habituation and recognition memory performance as predictors of later IQ. *Child Development* **64**(1), 57–79.

McShane, B. B., Gal, D., Gelman, A., Robert, C. & Tackett, J. L. (2019). Abandon statistical significance. *American Statistician* **73**(sup1), 235–245.

Meints, K., Plunkett, K. & Harris, P. L. (2008). Eating apples and houseplants: typicality constraints on thematic roles in early verb learning. *Language and Cognitive Processes* **23**(3), 434–463.

Melzer, A., Prinz, W. & Daum, M. M. (2012). Production and perception of contralateral reaching: a close link by 12 months of age. *Infant Behavior and Development* **35**(3), 570–579.

Meyerhoff, H. S., Papenmeier, F., Jahn, G. & Huff, M. (2013). A single unexpected change in target- but not distractor motion impairs multiple object tracking. *I-Perception* **4**(1), 81–83.

Michotte, A. (1946). *The perception of causality*. New York: Basic Books.

Mirman, D. (2014). *Growth curve analysis and visualization using R*. Boca Raton, FL: CRC Press.

Mirman, D., Dixon, J. A. & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: growth curves and individual differences. *Journal of Memory and Language* **59**(4), 475–494.

Nakagawa, S., Johnson, P. C. D. & Schielzeth, H. (2017). The coefficient of determination $R^2$ and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface* **14**(134), e20170213.

Nakagawa, S. & Schielzeth, H. (2013). A general and simple method for obtaining $R^2$ from generalized linear mixed-effects models. *Methods in Ecology and Evolution* **4**(2), 133–142.

Oakes, L. M. (1994). Development of infants' use of continuity cues in their perception of causality. *Developmental Psychology* **30**(6), 869–879.

Oakes, L. M. & Cohen, L. B. (1990). Infant perception of a causal event. *Cognitive Development* **5**(2), 193–207.

R Core Team. (2019). *R: a language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing.

Saussure, F. de (1916). *Course in general linguistics (Cours de linguistique générale)*. (C. Bally & A. Sechehaye, eds, R. Harris, trans.). La Salle, IL: Open Court Publishing.

Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods* **8**(2), 597–599.

Schlottmann, A., Ray, E. D. & Surian, L. (2012). Emerging perception of causality in action-and-reaction sequences from 4 to 6 months of age: Is it domain-specific? *Journal of Experimental Child Psychology* **112**(2), 208–230.

Shafto, C. L., Conway, C. M., Field, S. L. & Houston, D. M. (2012). Visual sequence learning in infancy: domain-general and domain-specific associations with language. *Infancy* **17**(3), 247–271.

Shimizu, Y. & Johnson, S. C. (2004). Infants' attribution of a goal to a morphologically unfamiliar agent. *Developmental Science* **7**(4), 425–430.

Singh, L., Steven Reznick, J. & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: a longitudinal analysis: infant word segmentation and vocabulary outcomes. *Developmental Science* **15**(4), 482–495.

Talmy, L. (1988). Force dynamics in language and cognition. *Cognitive Science* **12**(1), 49–100.

Tomasello, M. (2003). *Constructing a language: a usage-based theory of language acquisition*. Cambridge, MA: Harvard University Press.

Tomasello, M. & Vaish, A. (2013). Origins of human cooperation and morality. *Annual Review of Psychology* **64**(1), 231–255.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition* **69**(1), 1–34.