How do infants recognize joint attention?

Erik Gustafsson, Julie Brisson, Christelle Beaulieu, Marc Mainville, Dominique Mailloux, Sylvain Sirois

Abstract
The emergence of joint attention is still a matter of vigorous debate. It involves diverse hypotheses ranging from innate modules dedicated to intention reading to more neuro-constructivist approaches. The aim of this study was to assess whether 12-month-old infants are able to recognize a “joint attention” situation when observing such a social interaction. Using a violation-of-expectation paradigm, we habituated infants to a “joint attention” video and then compared their looking time durations between “divergent attention” videos and “joint attention” ones using a 2 (familiar or novel perceptual component) × 2 (familiar or novel conceptual component) factorial design. These results were enriched with measures of pupil dilation, which are considered to be reliable measures of cognitive load. Infants looked longer at test events that involved novel speaker and divergent attention but no changes in infants’ pupil dilation were observed in any conditions. Although looking time data suggest that infants may appreciate discrepancies from expectations related to joint attention behavior, in the absence of clear evidence from pupillometry, the results show no demonstration of understanding of joint attention, even at a tacit level. Our results suggest that infants may be sensitive to relevant perceptual variables in joint attention situations, which would help scaffold social cognitive development. This study supports a gradual, learning interpretation of how infants come to recognize, understand, and participate in joint attention.

1. Introduction

At around 6 months of age, infants begin to follow shift of gaze or head turn of adults, and by 12 months of age are able to actively coordinate focus on an object with a second person (Butterworth & Jarrett, 1991; Corkum & Moore, 1998; Morales et al., 2000; Scaife & Bruner, 1975). Labeled as “joint attention”, this ability for a triadic interaction between two people and an object has been described as a key component of our social cognition, allowing the sharing of experience and knowledge (Heal, 2005). In fact, infants’ skills in initiating and responding to joint attention predict their linguistic, social, and emotional abilities in later life (Morales et al., 2000; Mundy et al., 2007; Mundy & Gomes, 1998; Mundy & Newell, 2007; Parlade et al., 2009; Vaughan Van Hecke et al., 2007). For this reason, elucidating phylogenetic and ontogenetic factors trigger the emergence of joint attention is crucial for understanding the development of human social cognition. However, this is still a matter of vigorous debate (see Tomasello, Carpenter, Call, Behne, & Moll, 2005a).
Several non-exclusive hypotheses have been proposed, often based on the relationship between “joint attention” and the understanding of other persons as intentional agents (Tomasello, 1995). One line of hypothesis proposes hardwired modules dedicated to intention reading, such as an action-interpretation system that perceives action as goal directed (Gergely & Csibra, 2003), or a “shared-attention mechanism” involving innate modules perceiving goals and eye gaze direction (Baron-Cohen, 1997). Such modules may be involved in joint attention by helping the infant either to monitor the adult reaction when pointing to an object, or to become aware when the adult wants to share the perception of an object. Other studies focused on mirror-neuron systems believed to mediate action understanding in humans (Buccino, Binkofski, & Riggio, 2004; Fadiga, Fogassi, Pavesi, & Rizzolatti, 1995; Kohler et al., 2002), but whether such neurons are innate, or shaped by experience, is still debated (Ferrari, Tramacere, Simpson, & Iriki, 2013) as is their role in action understanding (Hickok, 2009). Presumably, the developmental pathway for understanding and responding to joint attention depends on early social interactions.

An alternative approach addresses the identification of such ontogenetic processes and their role. Dominant contemporary perspectives under the social-cognitive model suggest that infants begin to understand particular kinds of intentional states in others only after they have experienced them first in their own activity (Falck-Ytter, Gredebäck, & von Hofsten, 2006; Sommerville & Woodward, 2005; Tomasello et al., 2005a). According to this model, initiating and responding to joint attention should be highly correlated processes, insofar as they depend upon the same association between intention and goal-related behaviors in oneself and in others. The neuronal systems supporting such identification with others would flourish through certain typical sensorimotor and social contingencies such as protoconversation or mirroring behavior of parents toward their infant (Heyes, 2010; Trevarthen, Kokkinaki, & Fiamenghi, 1999). Epigenetic effects may also be involved, although the specific socio-environmental stimuli likely to trigger these molecular changes are yet to be identified (Ferrari et al., 2013). Moreover, one problem in addressing the potential ontogenetic processes implicated in understanding of joint attention in infants concerns the wide variety of child-rearing practices across human cultures. In fact, it is not known whether children across cultures are similarly exposed to supposedly typical social interactions such as protoconversation and mirroring (Bhavnagri, 1986; Bhavnagri & Gonzalez-Mena, 1997; Kagan & Klein, 1973; Ochs & Schieffelin, 2001). Considering the diversity of human ontogenetic niches, the mechanisms involved in ability to respond to communicative signals, such as those implicated in joint attention, would have to be extremely flexible.

Previous studies have largely focused on how infants manage to respond to joint attention. However, it is important to note that initiating joint attention and responding to it are abilities that should be considered separately (Mundy & Newell, 2007). Indeed, Mundy (2003) demonstrated that children diagnosed with autism, a disorder characterized by social interaction impairment,
show capacities for responding adequately to joint attention but rarely for initiating joint attention behavior. A similar pattern of results was found in a study investigating social cognition in chimpanzees (Tomasello, Carpenter, & Hobson, 2005b). Moreover, responding to joint attention can be measured as early as 6 months of age; that is, at least 3 months before infants can initiate a joint attention situation (Mundy & Newell, 2007). Taken together, these findings suggest that distinct mechanisms could underlie different features of joint attention abilities, suggesting that alternative models are required to fully explain the development of joint attention in infancy.

Mundy, Card, and Fox (2000) and Mundy, Sullivan, and Mastergeorge (2009) proposed such a model based on the fact that responding to joint attention activates the posterior neuronal systems dedicated to orientation and perceptual attention while initiating joint attention is related to the late maturing anterior attention systems in infants. Once adult, imaging data suggest an integrated activity of anterior and posterior systems during joint attention situations (Henderson, Yoder, Yale, & McDuffie, 2002; Williams, Waiter, Perra, Perrett, & Whiten, 2005). This attention-system model suggests that learning about and from self-control attention is the first step to joint attention. At a later stage, maturation of cortical networks, in conjunction with experiencing joint attention situations, would result in a sufficiently organized, accurate, and fast integrated processing of information between anterior and posterior neuronal networks thought to support intention reading and goal understanding (Mundy & Newell, 2007). Such a model represents a more parsimonious hypothesis compared to the social-cognitive one. Indeed, it suggests that before gaining an accurate comprehension of the conceptual features of joint attention supported by intention reading abilities, infants may just initiate and respond to joint attention based on orienting and attention systems elicited by gaze direction.

The aim of this study was to assess whether 12-month-old infants are able to recognize a “joint attention” situation between two people (one pointing out a toy to the other) as a specific social interaction. For that purpose, using a violation-of- expectation paradigm, we compared looking time durations between “divergent attention” and “joint attention” situations. Infants are expected to look longer in a “divergent” condition since the novelty of the situation should elicit stronger reactions from infants. However, preferential looking at a conceptually novel event is often confounded with that event’s perceptual novelty (Hunter & Ames, 1988; Jackson & Sirois, 2009; Roder, Bushnell, & Sasseville, 2000). In order to explore independently and jointly the effects of conceptual novelty and simple perceptual novelty, we used a 2 (familiar or novel component) × 2 (joint or divergent attention) factorial design (Bogartz, Shinskey, & Speaker, 1997). However, conceptually distinct events are, by necessity, perceptually distinct (Sirois & Mareschal, 2002), making looking time data potentially equivocal. For this reason, the substantial and relevant information provided by looking time durations will be enriched with measures of pupil dilation. Although primarily a function of luminance, pupil dilation is also an unbiased indicator of information processing load (Beatty,
Pupil diameter is positively associated with cognitive arousal and has already been used to investigate the detection of impossible events and irrational social interactions by infants (Gredebäck & Melinder, 2010; Jackson & Sirois, 2009; Karatekin, 2004, 2007; Porter, Troscianko, & Gilchrist, 2007; Sirois & Jackson, 2012). Rather than contrast cumulative looking time durations to sequences that are identical or similar apart from one key event or feature, change in pupil size can be measured concomitantly with the unfolding diverse events. Thus, if infants recognize the intention underlying joint attention, predictions for pupil diameter are that at key moments during the time-course of trials, pupils should dilate more for divergent attention than joint attention.

2. Method

2.1 Participants
30 infants (15 male, 15 female), healthy and full-term, were tested in the experiment. Mean age was 12 months, 11 days (SD = 0 months, 7 days). 11 additional infants participated but were excluded from analyses due to fussiness/crying, or for not looking during testing. Infants were recruited through poster adverts and direct information sessions to parents through a local baby resources center and hospital vaccination centers. Parents were from varied socioeconomic backgrounds. Once at the laboratory, parents provided written consent according to the guidelines specified by the Ethics Committee of the University du Québec à Trois-Rivières and the Regional Hospital Centre of Trois-Rivières.

2.2 Apparatus
During the experiment, infants sat on their parent’s lap in a dimly-lit cubicle. A video monitor (screen size 60 cm × 34 cm, 1920 × 1080 pixel) was placed on a table covered with a black cloth. The screen was 60 cm away from the baby. Speakers were located behind the monitor, on both sides. The experimental equipment was operated from outside of the cubicle. A camera mounted to the ceiling allowed the experimenter to monitor the infant and parent during the experiment. Gaze and pupil dilation were measured using a Tobii X120 eye tracker, at a sampling rate of 60 Hz. Stimulus presentation was carried out with E-prime 2.0 software and the E-prime extensions for Tobii software library.

2.3 Research design
We habituated infants to a short video sequence of joint attention in which two women are sitting at a table with three toys placed in front of them (a dinosaur, a football, and a fire truck). The women turn toward one another within about a second and conversation between the women starts at around 1 s (varied from 0.9 to 1.1 s) from the beginning of the video. The first speaker says: “My football is beautiful” while pointing to it.
The second speaker answers: “Yes” while turning her head toward the “joint” or the “divergent” toy at around 2 s (varied from 2.2 to 2.6 s).

First speaker: “Do you want to play with my football?” at around 3 s (varied from 3.1 to 3.7 s)

Second speaker: “Yes”, taking the football (“joint attention” condition) or the toy (dinosaur or fire truck) placed in front of her (“divergent attention” condition) at around 5 s (varied from 4.3 to 5.3 s).

Both women were wearing black shirts, had their hair tied back, had no make-up and did not wear glasses so that their faces and eyes were clearly visible. The actresses were two French-speaking Caucasian women. The background was light gray (see Fig. 1). Each video sequence, shown in 1920 × 1080 resolution, was played at 30 frames per second, with a duration of 7.75 s (225 frames). A still picture of the first frame of the video appeared at the beginning of each video sequence. The video started playing when the infant looked at the still picture for a minimum of 200 ms. Once the video ended, the last frame was frozen until the infant looked away for at least 2 s in total. Habituation was infant-controlled, with a maximum of 12 trials. We considered habituation to have occurred when we measured a 50% decrease in mean looking time on three consecutive trials, relative to mean looking on the first three trials (after Horowitz, Paden, Bhana, & Self, 1972).

Fig. 1. Screenshots from a “divergent attention” video, showing (a) the start of the event, (b) the first woman talking while pointing, (c) the second woman answering “yes” while looking at her toy and (d) the second woman taking her toy.

Next, the test trials began. Infants were shown four videos in random order, forming a 2 (first speaker, familiar or novel) × 2 (joint or divergent attention) factorial design. (1) The same video as the habituation phase (joint attention and familiar speaker), (2) a joint attention video with the
second woman showing the football (joint attention and novel speaker),
(3) a divergent attention video, wherein the second woman keeps looking at her own toy (the one placed in front of her) while answering, and then takes her toy (divergent attention and familiar speaker), and (4) the divergent attention video with the second woman showing the football (divergent attention and novel speaker).

The woman who speaks first was counterbalanced across participants. The same 4 video sequences were used with all infants. We measured looking time duration as well as pupil dilation throughout the experiment.

2.4 Procedure
The infant sat on their parent's lap in the cubicle, in front of the monitor. They were positioned so that the eye tracker could successfully capture the infant's gaze. Parents were told not to interact with their infant during the experiment. After closing the door, a five-point calibration procedure was performed following which, the experiment began. Infants were shown the habituation trials first. Then, they were shown the four test trials in a random order, for which pupil dilation and gaze data were collected. After the procedure, infants and parents were thanked and given a diploma and a reusable shopping bag as compensation for their participation.

2.5 Analyses
We measured infants' looking time duration to the still image at the end of each video in order to remove the intrinsic attractive effect of the video clip, to focus only on the attention holding it provokes (Woodward, 1998, 1999; Wertz and Wynn, 2014). A 2 (speaker) by 2 (attention) repeated-measures ANOVA was performed on the data.

The four videos varied in the timing of the key events in the sequence (i.e., pointing, head turn, and toy taking). For the analyses of pupil dilation, the average times of these events were used to warp data samples associated with each video to a common time sequence. Data segments around those key times were compressed or stretched, as required, to create aligned sequences of identical lengths. Pupil diameter data were recorded for both eyes at a sampling rate of 60 Hz throughout trials. Following the method of Jackson and Sirois (2009), we transformed the raw data from the test trials into smooth curves, baseline corrected by subtracting the pupil dilation in the sample preceding the video display. We then performed the 2 (speaker) by 2 (attention) repeated-measures ANOVA on the curves themselves, using Functional Data Analysis (FDA; Ramsay & Silverman, 1997).

Presenting all test conditions to each infant may have masked otherwise noticeable pupil-dilation results. For instance, after seeing a single divergent attention trial infants may show less interest in
further divergent trials. To control for an effect of trial order, we also ran a 2 (speaker) by 2 (attention) ANOVA for looking time duration on the first test trial, and analyzed pupil dilation on that first test trial.

We removed three individuals from the analysis because of floor effects and missing data.

3. Results

3.1 Looking time analyses

The average duration of looking over the first three familiarization trials across infants was 5.49 s. Of the 30 infants in the sample, only one failed to show a 50% decrease on its last three familiarization trials. In the test trials, the interaction between “attention” and “speaker” for looking times did not reach significance ($F(1,29) = 0.085, p = 0.773$, partial $\eta^2 = 0.003$). There was a significant main effect of “attention”, ($F(1,29) = 4.588, p = 0.041$, partial $\eta^2 = 0.137$), and infants tended to look longer when speaker roles were novel, relative to habituation trials ($F(1,29) = 3.396, p = 0.076$, partial $\eta^2 = 0.105$) (Fig. 2).

Fig. 2. Mean looking times (±SEM) of test trials as a function of attention and speaker.
When considering only the first test trial, the interaction between “attention” and “speaker” did not reach significance ($F(1,29) = 0.868, p = 0.360, \text{partial } \eta^2 = 0.032$). There was no significant effect of “attention”, ($F(1,29) = 1.333, p = 0.259, \text{partial } \eta^2 = 0.049$) or speaker ($F(1,29) = 0.239, p = 0.629, \text{partial } \eta^2 = 0.009$) (Fig. 3).

3.2 Pupil diameter analyses

Raw pupil data for each trial were resampled according to the procedure outlined in Section 2, so that key events were aligned in time across trials. The data were then transformed into curves ($b$-splines), following the methods outlined in Jackson and Sirois (2009), see also Sirois and Jackson (2012). Mean curves for each type of test trial are shown in Fig. 4. These are 4th order polynomials, with 18 bases.
Statistical analyses can be performed on the curves using a set of procedures from FDA (Ramsay & Silverman, 1997). Unlike the standard ANOVA, which only produces a discrete value (the $F$-ratio) as the outcome of the analysis, functional ANOVA produces a curve as its outcome, which can be plotted over time to assess if and, crucially, when interaction or main effects are observed. Readers are referred to Jackson and Sirois (2009) for a detailed description. Fig. 5 plots the outcome of a repeated-measures functional ANOVA on pupil diameter, with attention (joint or divergent) and speaker (familiar or novel) as repeated factors. None of the curves exceed the critical value of the relevant $F$ distribution (shown as a dashed horizontal line in the figure).
Mean curves for the first test trials are shown in Fig. 6. Figs. 7 and 8 plot the mean curves and the outcome of an independent t-test pupil diameter with attention (joint or divergent) and speaker (familiar or novel) as independent variables. When considering only the first test trial, both curves exceed the critical value of the relevant t-test distribution (shown as a dashed horizontal line in the figure). We observed a significant effect of the divergent condition on pupil dilation at the very beginning of the sequence, and from when the person looked at the wrong toy until the end of the video (Fig. 7).

**Fig. 6.** Mean b-splines for the first test trials. Dotted vertical lines indicate key moments (in order: 1st speaker pointing, 2nd speaker looking at the toy, 1st speaker talking, 2nd speaker taking the toy).

**Fig. 7.** (a) Mean b-splines for the first test trials involving “Divergent attention” and “Joint attention”. Dotted vertical lines indicate the key moments as in Fig. 3. (b) Main effects from the independent t test. The dashed horizontal lines are the critical t-values.
Novel speaker elicited higher pupil dilation at the very beginning of the video, the first time the novel speaker says the otherwise familiar utterance. However, novel speaker elicited lower pupil dilation thereafter (Fig. 8).

4. Discussion

Infants tended to look longer at test events that involved novel speaker and significantly longer when events involved divergent attention. In the absence of a significant interaction, these are considered additive, independent effects. The longer looking duration suggests that infants perceive the change in speaker, and the “divergent” attention. This raises the question of whether the longer gaze observed in the “divergent” condition is due to the novel perceptual features (person taking another toy) or to the novel conceptual features of this condition (incongruence with the “laws of communication”). No significant changes in infants’ pupil dilation were observed during the “novel speaker” or the “divergent attention” conditions when considering all test trials.

When focusing only on the first test trial, we did not find significant effects regarding looking time duration. This may be due to the sample size, which is reduced as a function of isolating only the first of the four test trials. However, significant effects were observed for pupil dilation. We noted greater pupil dilation when the novel speaker starts saying the familiar utterance, and when speakers begin to look at the “wrong” toy. Interestingly, pupil dilation was significantly lower the second time the novel speaker said the familiar utterance, suggesting such novelty provoked a lower attention or a lack of interest after the initial surprise. The fact that divergent attention elicited higher pupil dilation when the speaker looked at and took the “wrong” toy suggests that a violation of infants’ expectations occurred. Nonetheless, our results cannot definitively answer the question of whether “intention reading” is present in young infants, nor if it is required to recognize “joint attention” situations. Looking time data and pupillometry suggest that infants may well appreciate discrepancies from expectations related to joint attention behavior. However, since significant pupil
dilation was also elicited from presumed perceptual changes (speaker’s role), the present results suggest that infants may be especially sensitive to relevant perceptual variables in joint attention situations, which would help scaffold social cognitive development. Rather than speculating in favour of intention reading innate modules, or assuming that infants may understand intentional actions (Falck-Ytter et al., 2006; Tomasello et al., 2005a), this study supports an alternative, more parsimonious interpretation. As proposed by Bogartz et al. (1997), by comparing what is observed with the information in memory, infants develop their associative memory and eventually some prediction or expectation regarding others’ actions. A discrepancy between what is perceived and the information in memory would result in attention holding (Cohen, 1972). Without denying a possible role of innate modules or mirror neuron systems, our results highlight the importance of first considering the role of already well-accepted perceptual and learning mechanisms, that are equally suited to conveying information to memory and constructing immediate and long-term representations. This interpretation is consistent with the attention model systems proposed by Mundy et al. (2000), as well as connectionist and constructivist models suggesting that only progressive integration of distributed cortical networks through bi-directional interactions can sustain the development of self-awareness levels necessary for intention-reading and conceptual understanding (Decety & Sommerville, 2003; Keysers & Perrett, 2004; see also Piaget & Cook, 1952).

4.1 Limitations and perspectives
Looking at the first test trial in isolation both reduces statistical power and increases the risk of artefactual findings. This might account for the significant effect elicited by the novel speaker before the first second, despite there being no visible difference between conditions at that time. While results from some analyses may inform future research, they also highlight that the effects are statistically weak when they do not show up in the full analysis, which is statistically more powerful (especially when using, as we have, a repeated-measures design). This further suggests caution when attributing complex attributions in infants. To conclude, in light of the diversity of child-rearing practices documented within the field of anthropology, complex cognitive skills which can emerge from simple learning processes appear more adaptive than skills requiring complex species-specific socio-environmental stimuli to develop. Accordingly, adopting a more conservative stance concerning infants’ actions permits a focus on basic mechanisms that may underlie the emergence of higher order cognitive skills and complex behaviors in humans (Provine, 2005; Chittka, Rossiter, Skorupski, & Fernando, 2012).
Acknowledgements

We thank the parents who kindly agreed to participate in this study, and their babies. We are grateful to the research assistants (Lisa Auger, Vanessa Bergeron, Geneviève Lessard, Sara-Valérie Giroux, Isabelle Gosselin, Josianne Aylwin) who welcomed the parents and helped with the experimental procedures. We also thank Alisha Gibson for reviewing an earlier version of the manuscript. This work received funding from the Canada Research Chair (NSERC) program (grant 950-219744) and the Social Sciences and Humanities Research Council of Canada (grant 410-2010-2125).

References


