Measuring the development of social attention using free-viewing

Michael C. Frank

Department of Psychology, Stanford University

Edward Vul

Department of Psychology, University of California, San Diego

Rebecca Saxe

Department of Brain and Cognitive Sciences, MIT

We gratefully acknowledge the parents and children who participated in this research; the staff of the Boston Children's Museum for their generous accommodation of our research; Laura Schulz, Darlene Ferranti, and Ali Horowitz for help facilitating data collection at the Children's Museum; Dima Amso, Tom Fritzche, Scott Johnson, and Tamar Kushnir for helpful feedback and discussion; and Allison Gofman, Erica Griffith, Avril Kenney, and Arathi Ramachandran for help in data collection. This research was supported by a Jacob Javits Graduate Fellowship and NSF DDRIG #0746251.

Please address correspondence to Michael C. Frank, Department of Psychology, Stanford University, 450 Serra Mall, Jordan Hall (Building 420), Stanford, CA 94309, tel: (650) 724-4003, email: `mcfrank@stanford.edu`.

Abstract

How do young children direct their attention to other people in the natural world? While many studies have examined the perception of faces and of goal-directed actions, relatively little work has focused on what children will look at in complex and unconstrained viewing environments. To address this question we showed videos of objects, faces, children playing with toys, and complex social scenes to a large sample of infants and toddlers between 3 and 30 months old. We found systematic developmental changes in what children looked at. When viewing faces alone, younger children looked more at eyes and older children more at mouths, especially when the faces were making expressions or talking. In the more complex videos, older children looked more at hands than younger children, especially when the hands were performing actions. Our results suggest that as children develop they become better able to direct their attention to the parts of complex scenes that are most interesting socially.

How do young children see other people, and what aspects of others do they focus on? Social attention—defined here as the process by which observers select and encode aspects of other people—has been studied extensively from several different perspectives. Research on this topic has examined the development of face perception, the perception of goal-directed action, person-detection, and many other aspects of social attention (reviewed in e.g. Nelson, 2001; Gergely & Csibra, 2003; Gredebäck, Johnson, & Von Hofsten, 2010). But despite the prominence of these lines of work, relatively little research has examined what is arguably the most direct measure of social attention: what children choose to look at in unconstrained displays. The current study uses free-viewing eye-tracking to assess social attention in complex natural scenes at a wider range of ages than has previously been studied and across a variety of different social contexts. Our goal is to understand what kinds of social information infants and children seek out in complex scenes and how the use of this information changes across development.

A wealth of research has examined infants' and young children's perception of faces and goal-directed actions. This work has largely used schematic or photographic displays in isolation to make a controlled assessment of preference or discrimination. Results from this work suggest that newborn infants prefer faces to matched stimuli (Farroni et al., 2005; Johnson, Dziurawiec, Ellis, & Morton, 1991; Simion, Cassia, Turati, & Valenza, 2001; Morton & Johnson, 1991) and that over the course of the next several months, infants gain the ability to make finer distinctions between identities (Pascalis, De Haan, Nelson, & De Schonen, 1998), genders (Quinn, Yahr, Kuhn, Slater, & Pascalis, 2002), and faces of their own race (Kelly et al., 2005) and species (Pascalis, Haan, & Nelson, 2002). Infants also are able to encode the goal of a reach by six months (Woodward, 1998) and only a few months later they are quite sophisticated at inferring the intentions underlying a gesture (Yoon, Johnson, & Csibra, 2008) or the motion of a geometrical shape (Csibra, Gergely, Bıro, Koos, & Brockbank, 1999; Gergely, Nádasdy, Csibra, & Bíró, 1995). They

are even able to infer the goal of an action when that action is not completed (Brandone & Wellman, 2009; Hamlin, Hallinan, & Woodward, 2008; Meltzoff, 1995). Thus, within the first year, young children have both a complex representation of faces and a sophisticated understanding of others' actions.

Because of these robust findings, researchers have begun to use social attention as a measure of group differences. For example, studies of infants and adults at risk for or diagnosed with Autism Spectrum Disorders have found that differential fixation of mouths over eyes in static and moving faces may be associated with autism (Dalton et al., 2005; Klin, Jones, Schultz, Volkmar, & Cohen, 2002; Merin, Young, Ozonoff, & Rogers, 2007). In addition, cross-cultural work has investigated differences in face looking patterns between European and East Asian adults (Blais, Jack, Scheepers, Fiset, & Caldara, 2008; Jack, Blais, Scheepers, Schyns, & Caldara, 2009). Social attention and its development is thus a central issue for researchers working in a wide variety of fields. Yet, perhaps due to methodological issues, relatively little work has examined the development of social attention in naturalistic displays that reflect the complexity of real world social interactions.

Two methods are beginning to be used to correct this imbalance. First, recent exciting work using a head-mounted camera has begun to map out the structure of infants' first-person visual experience. Yoshida and Smith (2008) explored the use of a head-mounted camera for recording the natural field of view of infants in free-play with a parent. They found that, compared with a 3rd person view, the child's visual experience (as captured by the head camera) was much more focused on one or a small set of objects and that it was far more likely to contain the child's own hands or the parent's hands as opposed to the parent's face. Aslin (2009) similarly gathered naturalistic recordings with a head-mounted camera, but then recorded eye-tracking data while showing these videos to 4- and 8-month-old infants and adults. They found differences in fixation across a variety

of activities (e.g. shopping elicited less looking at hands than play with blocks at home), and adults looked significantly more at the people than infants did. An in-depth analysis of these same stimuli also examined the motion properties of stimuli in the child's field of view (Cicchino, Aslin, & Rakison, 2010). While work with head-mounted cameras is still new, this method has tremendous potential to allow for detailed analyses of what children see and how their visual experience changes across development.

Second, there is a growing body of work using corneal-reflection eye-tracking to understand infants' viewing patterns in social situations. For example, infants' understanding of goal-directed actions, which can be measured using habituation paradigms (Woodward, 1998; Yoon et al., 2008), has also been probed via anticipatory eye-movements. Falck-Ytter, Gredebäck, and Von Hofsten (2006) showed that 12-month-olds and adults looked at the goal of an action (e.g. putting something in a bucket), whereas 6-month-olds did not; this finding has since been replicated and extended to suggest sensitivity to the particular action types being used (Gredebäck, Stasiewicz, Falck-Ytter, Rosander, & Hofsten, 2009). Similarly, infants' gaze-following behavior, which has typically been manually coded from live interactions (Scaife & Bruner, 1975), has been studied using eye-tracking methods in both controlled video displays (Hofsten, Dahlström, & Fredriksson, 2005; Gredebäck, Theuring, Hauf, & Kenward, 2008; Senju & Csibra, 2008) and even via eye-tracking of live interactions (Gredeb ack, Fikke, & Melinder, in press). The use of eye-tracking methods in both of these cases allows for the testing of participants in multiple conditions as well as a far greater degree of precision in measuring participants' reaction times on a trial-to-trial basis.

Although the displays used in much of this work are much more naturalistic than those used by previous studies, they still contain extensive repetition of individual actions (e.g. reaching or looking behavior) in order to collected repeated measurements of participants' responses and reaction times. Our recent work takes a slightly different

approach: we showed 3-, 6-, and 9-month-old infants (as well as a control population of adults) a narrative video stimulus—*A Charlie Brown Christmas*, an engaging, animated cartoon containing social interactions and complex backgrounds. We then used a variety of analytic methods to extract consistent behaviors (such as looking at faces or looking at perceptually salient regions) from the continuous eye-tracking data we collected (Frank, Vul, & Johnson, 2009). We found that, although all the groups in the study looked at faces, there was still a considerable increase in the amount that older infants and adults looked at faces relative to the youngest group (consistent with the difference observed by Aslin, 2009).

Our current study follows on these previous studies in addressing questions about social attention in complex and rich natural scenes. We were particularly interested in whether there were developmental changes in social attention beyond the first year and differences in how social attention was allocated depending on both the visual and social complexity of the context. Accordingly, we designed our recruitment procedure to include older infants and toddlers as participants. As our stimuli we chose a set of live-action movies of children playing accompanied by uncoordinated classical music (adapted from the Baby Einstein series). These stimuli eliminated intermodal regularities that could act as a confound in measuring social action (Klin, Lin, Gorrindo, Ramsay, & Jones, 2009). We systematically varied the amount of detail and complexity in our stimuli, breaking social stimulus videos into three conditions: (a) children's faces on a white background; (b) children playing with objects on a white background; and (c) multiple children playing, often with adults, in a real-world setting. As a control for developmental differences in visual complexity and motion processing, we included a set of videos of objects moving on a white or black background.

Together these materials allow for the examination of how the patterns found in previous stimuli—in particular, looking at faces and looking at hands during

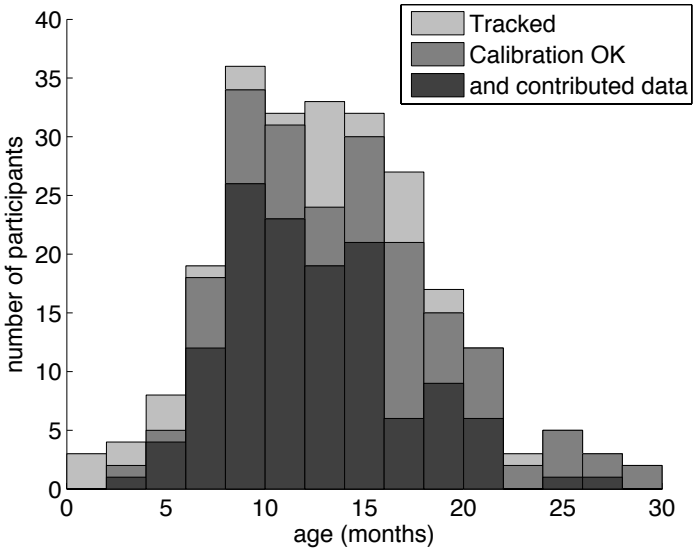

Figure 1. A histogram of the ages of our participants. The light gray histogram includes all participants; medium gray represents the subset those participants whose calibration could be verified and adjusted offline; and dark gray represents the subset who were included in the final sample: their calibration was acceptable and they contributed usable eye-tracking data for more than 20% of the stimuli.

actions—generalize across a wide range of ages and stimulus complexities, in the absence of important but confounding intermodal regularities.

Methods

Participants

Our recruiting followed an opportunistic design. Two hundred and thirty six children between the ages of 3 and 30 months were recruited from the PlaySpace (an area for children less than three years old to play freely) of the Boston Children's Museum via conversations with their parents during the course of a normal visit. Of those 236 children whose parents consented for them to participate in the study, we included data from 204

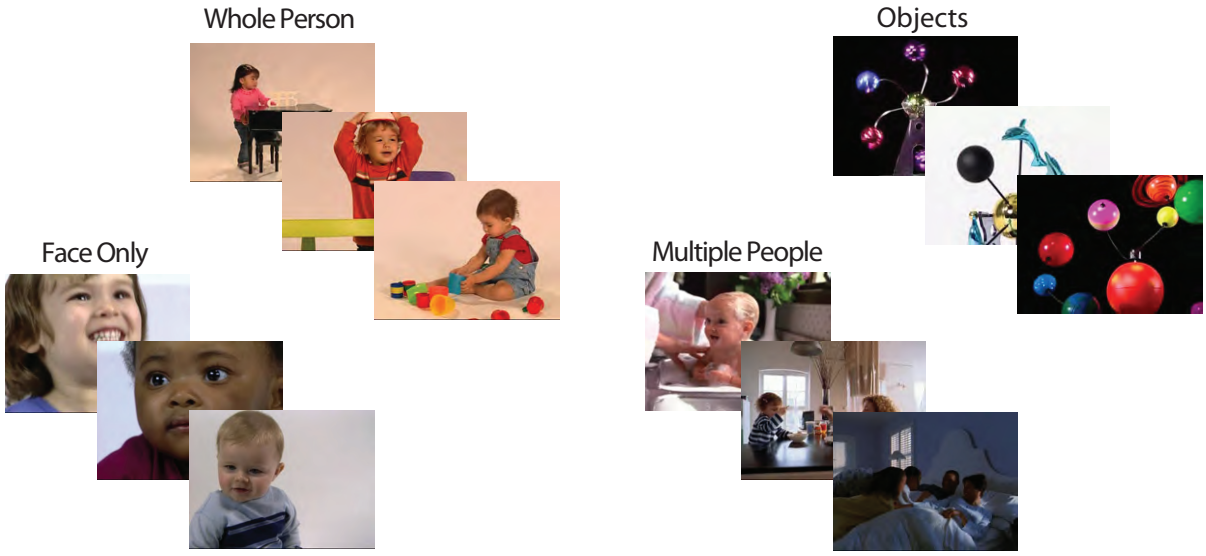

Figure 2. Three representative frames from the first video of each of the three social stimulus conditions and the object control condition.

(86.4%) who had calibrations that could be verified or adjusted offline (see below for details). Of those 204, 129 (63.2%) contributed eye-tracking data for more than 20% (48s) of the entire 240s main stimulus. We excluded children who contributed limited amounts of data because the sparsity of their data meant that we were not able to measure their behavior appropriately across conditions and videos—in many cases these children left the study after viewing the first video (often due to fussing or squirming). The 129 children who fulfilled our criteria for calibration and contribution of data constituted our final sample (mean age = 12.5 months, min = 3.2 months, max = 27.8 months). Figure 1 shows the age distribution of our sample before and after inclusion criteria were applied.

Stimuli

All stimuli were short, live-action videos accompanied by unsynchronized classical background music. Stimuli were constructed from four Baby Einstein videos (Walt Disney

Productions, 2002), a series of widely-available videos developed for infants and toddlers: *Baby Galileo: Discovering Sky*, *Baby Neptune: Discovering Water*, *Baby Monet: Discovering the Seasons* and *Baby Van Gogh: World of Colors*.

The stimulus set consisted of three 20s videos in each of four conditions. The four conditions were *Face Only*, *Whole Person*, *Multiple People*, and *Objects*. For each condition, we extracted short segments from the source videos while maintaining the soundtrack from a single video (for consistency). In the *Faces Only* condition, movies consisted of close-ups of children's faces (and occasionally their torsos and upper bodies), on a white or neutral background. The movies in the *Whole Person* condition included single children (now pictured in full) playing with toys on a white background, e.g. a toddler playing with a set of colored cups. The movies in the *Multiple People* condition included one or multiple children playing (often with adults) in normal indoor and outdoor settings, e.g. a mother and son eating breakfast. The *Objects* condition included videos of balls rolling around a track, colored mobiles, and other moving toys. Each 20s video consisted of between 4 and 7 clips consisting of a single camera shot with no cuts (min length = 1.67s, max length = 8.03s). Example frames from each condition are shown in Figure 2.

Also included in the stimulus set were three instances of an 11s calibration verification stimulus, which consisted of an image of a yellow toy star moving on a black background. The star moved to four different locations distributed around the screen, accompanied by a coordinated sound. This movie was shown at the beginning, midpoint, and end of the experiment.

Procedure

After giving informed consent, parents and children were escorted to a small room adjacent to the recruitment site. Children sat on parents' laps approximately 60cm away

from the monitor of a Tobii T60 binocular corneal-reflection eye-tracker. The monitor was mounted on an ergonomic arm to allow it to be adjusted to the height and angle of the child. The room was normally lit with diffuse fluorescent light from above. Parents were asked not to talk to or to try and influence their children in any way during stimulus presentation (but were not prevented from watching the videos themselves).

We first carried out the Tobii tracker's calibration routine using a two-point calibration and then immediately began showing the video stimuli. All stimuli were presented using Tobii Studio (the Tobii eye-tracker's proprietary software). Videos were presented in one of two random orders. We created each order by randomizing video order within three blocks, which each contained a single video from each of the four conditions (with the further constraint that no two videos from a singe condition were adjacent). The total duration of the experiment was approximately 4m 30s.

Data Preprocessing

All preprocessing and analyses were conducted with custom Matlab software unless otherwise specified. We first exported data from Tobii Studio. Since the Tobii tracker collects binocular data, we averaged across eyes, interpolating from a single eye when validity of the other was low. We next smoothed the tracked data using an adapted bilateral filtering algorithm (Durand & Dorsey, 2002; Frank et al., 2009). The purpose of this algorithm was to smooth out local variations in fixation due to tracker noise while retaining the magnitude and timing of saccadic changes in gaze position.

We next attempted to verify the precision of the calibration for each of our participants. Because we were interested in the development of looking at precise regions of interest, ensuring the accuracy of our data was very important to our conclusions. Without some external test of calibration accuracy, it could be the case that any developmental change we observed was caused by differences in the accuracy of calibration

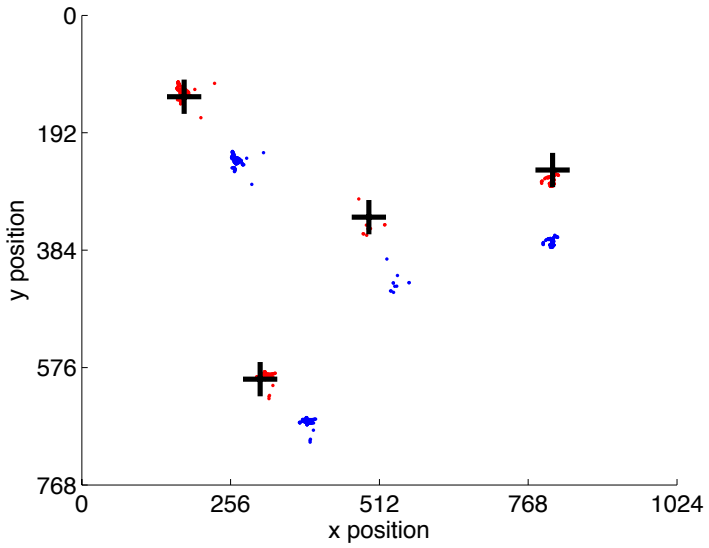

Figure 3. An example of a single child's eye-track on the four-point offline calibration stimulus we used. Blue dots represent the child's original point-of-gaze at each time point during the stimulus (excluding transitions between target locations), while red dots indicate point-of-gaze after adjustment. Black crosses represent the center position of the calibration object.

across ages. This concern motivated the inclusion of our "calibration check" stimulus in the experiment so that we could then use the position of participants' point-of-gaze during this stimulus as a ground-truth measurement for assessing accuracy.

Examining the records of individual infants' point-of-gaze, we discovered systematic errors in calibration (for an example, see Figure 3, blue points). We designed a procedure to correct this issue. We first isolated sections of children's track corresponding to the points at which the calibration stimulus was static (offset by 500ms to correct for delays in locating the target by the younger children in our sample). We then conducted parallel robust regressions (a method of regression which downweights points considered to be outliers, Holland & Welsch, 1977) in the X and Y planes to find the best translation and

expansion/contraction of the data to match the calibration points (Figure 3, red points). We then re-calibrated individual infants' track on the basis of these values.

We examined each infant's adjusted calibration by hand. We included participants for which there were a minimum of two adjustment points for which there was sufficient track and for which some part of the adjusted point-of-gaze made contact with the stimulus and excluded those infants for which the procedure had failed (either because there were not enough data or because fixations were scattered in ways that did not correspond to the calibration check stimulus). This exclusion is reflected in the sample description reported above. This procedure ensured a high degree of accuracy in the calibrations of those participants included in the study.[1]

Analysis methods

We created regions-of-interest (ROIs) for each video by using custom software to hand-code the bounding rectangle around stimuli of interest in each frame. For the Face Only condition, we coded faces, eyes, and mouths; for the Whole Person and Multiple People conditions, we coded faces and hands. (We assumed that even using our adjustment procedure, the margin of error by the tracker was likely too large to warrant coding eyes and mouths in the faces of the Whole Person and Multiple People conditions). In order to include eye-movements to the edges of particular ROIs (Haith, Bergman, & Moore, 1977) and to account for small deviations in calibration that remained after adjustment, we smoothed each ROI with a 15 pixel radius (approximately .5 degrees of visual angle). Modification of this parameter did not qualitatively alter the pattern of results for any analysis. For each child we extracted percent dwell-time within the coded ROIs for each condition. To avoid problems with sparse, noisy measurements of individual

[1]Code and further documentation for this procedure available at `http://langcog.stanford.edu/materials/calib.html`.

children, we excluded children from a particular condition if they did not contribute data from at least 18s (30%) of the total 1m video for that condition (note that this condition-by-condition exclusion criterion is applied only to data from those children who passed the separate subject-level exclusion criteria). When two regions of interest overlapped on a particular fixation, that fixation was counted as belonging to both regions.

For subsidiary content analyses, we divided video clips in the Faces Only condition into two action-groups: those which included mouth movements from talking or other related vocalization (without sound) or some facial expression involving mouth movements like smiling (11 clips)[2] and those which did not 7 clips). For clips in the Whole Person and Multiple People conditions, we divided the videos in this condition into three categories on the basis of how people in the videos used their hands: those in which the children in the videos used their hands only for holding or supporting actions (7 clips); those in which hands were used for picking up an object, putting down an object, or otherwise changing its position (11 clips); and those in which children used their hands for a more complex action (e.g. pointing, pouring, or banging on the keys of a piano) (13 clips). We then split the ROI data by action-group clip, using average looking at an ROI for each clip as our dependent measure.

For our statistical analyses, we used linear mixed-effects models (Gelman & Hill, 2007) using the `lme4` package in R (R Development Core Team, 2005) to quantify the effects of age, action-group (no mouth expression vs. talking/smiling), and ROI (eyes and mouths) on dwell-time.[3] Because average dwell times were distributed in a roughly

[2]Only three clips appeared to include mouth movements from vocalizations (children saying "boo," "bye," and yelling, respectively)—without sounds, of course—so we included these with other mouth-related facial expressions rather than analyzing them separately.

[3]Several features of our data made linear mixed-effects models preferable to—as well as more conservative than—standard ANOVA analyses. First, the crossed design of our data (with multiple observations for each participant and for each video clip) cannot be captured in an ANOVA framework. To control for these, we

exponential pattern, we used logit transforms to create a dependent measure that was normally distributed and hence appropriate for a linear model. After the logit transform was applied, we standardized the units of dwell-time by converting them to z-scores (we performed this step in order to increase the interpretability of coefficients). All p-values and confidence intervals reported in mixed-model analyses were derived from posterior simulation using the `languageR` package (Baayen, 2008); these p values represent the proportion of samples from the model's posterior probability distribution for which the β weight was in the opposite direction. This number can be interpreted as the probability of an error in the direction of a particular effect and, like a standard p-value, can be used to assess statistical significance.

For our analysis of fixation predictability, we created fixation probability maps for each group of participants, as in Frank et al. (2009). These maps were created by collecting each participants' fixations for each frame of the stimulus and then convolving these fixations with a Gaussian kernel (Hastie, Tibshirani, Friedman, & Franklin, 2005). The kernel we chose extended forward but not backward in time (indicating some probability of looking at the same spot soon after a participant had looked there) and symmetrically in space around the point of gaze. In practice, we chose a kernel with a standard deviation of 40 pixels and a temporal standard deviation of 33ms (though the results reported here were qualitatively similar for other parameter choices). We then split each group's probability maps for each video into their component clips, creating a set of between 16 and 21 separate maps for each condition.

---

included intercept terms ("random effects") of both participant and video clip in our models. Second, due to the limitations of eye-tracking, our dataset contained a number of missing trials, which also cannot be incorporated into an ANOVA framework. In all cases, standard ANOVA analyses give qualitatively similar results, but the levels of significance and effect sizes from the mixed model analyses are more reliable than those from ANOVA so we do not report the ANOVA results here.

We quantified the predictability of fixations within each probability map by computing the entropy of that map. Entropy is an information-theoretic measure of the uncertainty within a probability distribution that gives the number of bits necessary on average to describe a sample from that distribution (MacKay, 2003). A larger number of bits corresponds to greater uncertainty about where a sample from the distribution will come from; in our study, larger entropy values map on to a larger spread of fixation (and hence less predictability). Because entropy is defined only over probability distributions (not individual observations or probabilities), each probability map yielded a single measurement of entropy, resulting in a set of entropy measurements for each group for each clip.

Results

Our goal was to measure the distribution of children's fixations to social regions of the stimuli across development. We began with region-of-interest analyses, examining which aspects of the stimuli were fixated at different ages. We next performed finer-grained analyses that divided the stimuli by their content. Finally, we performed a control analysis that examined whether the predictability of children's fixations across the stimulus materials differed across ages.

Region-of-interest analyses

We observed large developmental changes in the distribution of children's looking. Figure 4 shows individual participants' looking at the ROIs we coded for each of the three stimulus conditions. Because our naturalistic stimuli differed from one another on many dimensions, ROIs could only be compared within an individual stimulus, not across stimuli, thus the important trends shown in Figure 4 are the developmental trends in each individual subplot. Supplementary videos S1–S3 show ROIs and fixations for one movie

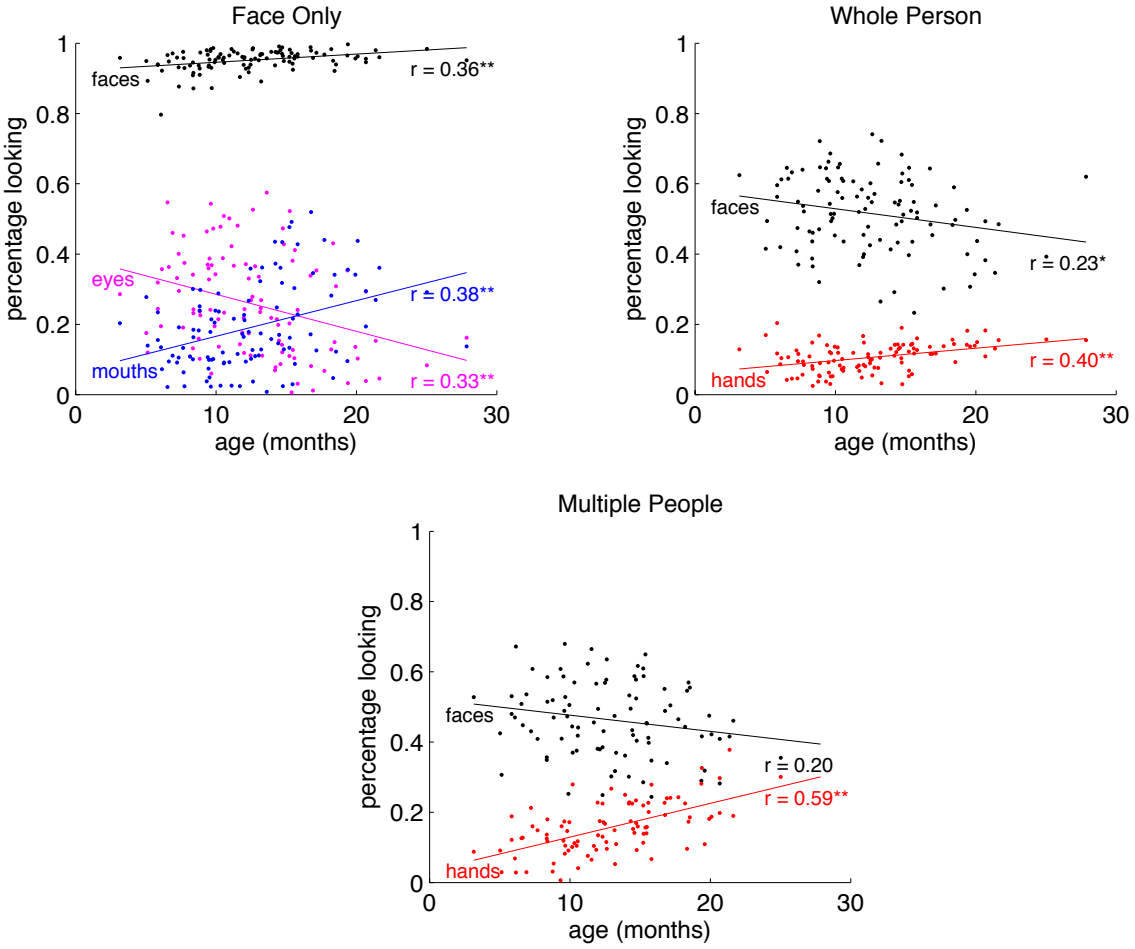

Figure 4. Each panel shows participants' percentage looking to the regions-of-interest that we coded for a particular condition, plotted by their age. Lines represent standard regression lines; r values and significance values are derived from these regressions (* = $p < .05$, ** = $p < .01$).

from each of the three conditions (available at `http://langcog.stanford.edu/materials/social_attention.html`). This analysis was not performed on data from the Objects condition. We did not believe that it was appropriate for several reasons: first, what constitutes an object (as opposed to the background, which is often also composed of objects) is often a subjective judgment; second, because of the quick movement in the videos (e.g., pendulums swinging, drops of oil falling, or billiard balls zooming around a track), even pilot adult participants could not track exactly and instead fixated parts of the objects' trajectory.

In the Face Only condition, we saw an intriguing developmental flip: younger children spent more time looking at eyes ($r = -.33$, $p = .0005$) and older children spent more time looking at mouths ($r = .38$, $p = .0001$). More generally, nearly all fixation time was spent looking at faces (95.2%). This ceiling effect was unsurprising because on average 59.7% of the total area of the movie was filled by the face ROI in this condition and the background was largely blank. Nevertheless, there was still a small but significant increase in looking at faces across development ($r = .35$, $p = .0002$).

In the Whole Person and Multiple People conditions, we observed a highly consistent increase in looking to hands for older children in both conditions ($r = .40$, $p < .0001$ and $r = .59$, $p < .0001$). In addition, we saw less overall looking at faces in these conditions compared with the Face Only condition (51.5% and 46.3%, respectively), though faces made up much less of the overall area of the movie (7.2% and 10.3%, respectively). Also in contrast to the Face Only condition, in both conditions containing more complex actions we saw a slight overall decrease in looking to faces with age ($r = -.22$, $p = .02$ and $r = -.20$, $p = .06$).

These results mirror important developmental changes over the period we studied—including increasing understanding of goal-directed action and increased knowledge of language and others' emotions—and suggest that preferences to attend to

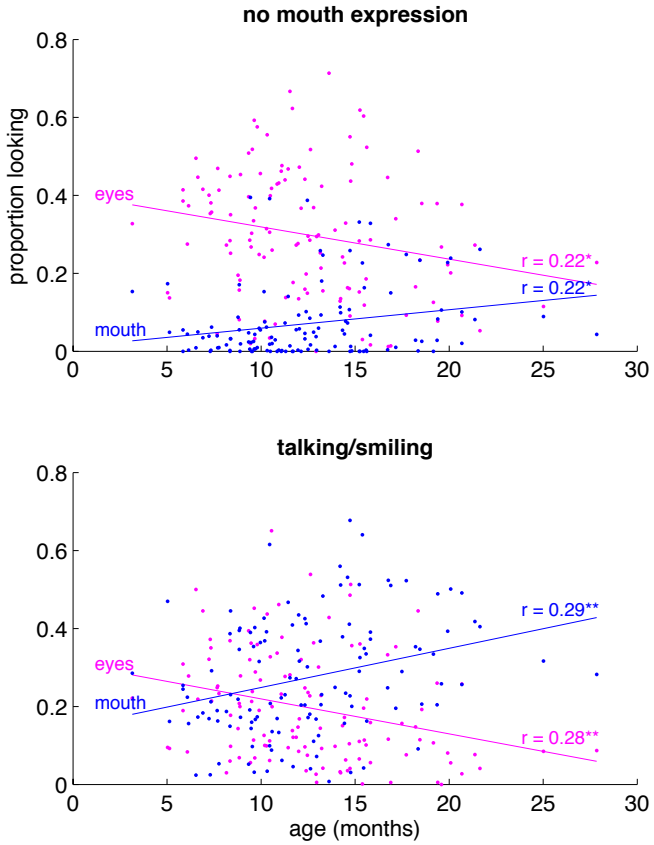

Figure 5. Each panel shows the proportion looking at eyes and mouths plotted by participants' ages within a subset of the clips in the Face Only condition. Plotting conventions are as in Figure 4.

individual regions like eyes or hands are not static across development.

Content analyses

In our next set of analyses we followed up on the individual ROI analyses by examining whether the social content of video clips had influenced our participants' looking behavior. We found that social content had a large impact on where children looked. Children looked more at mouths than at eyes when the mouths were talking or

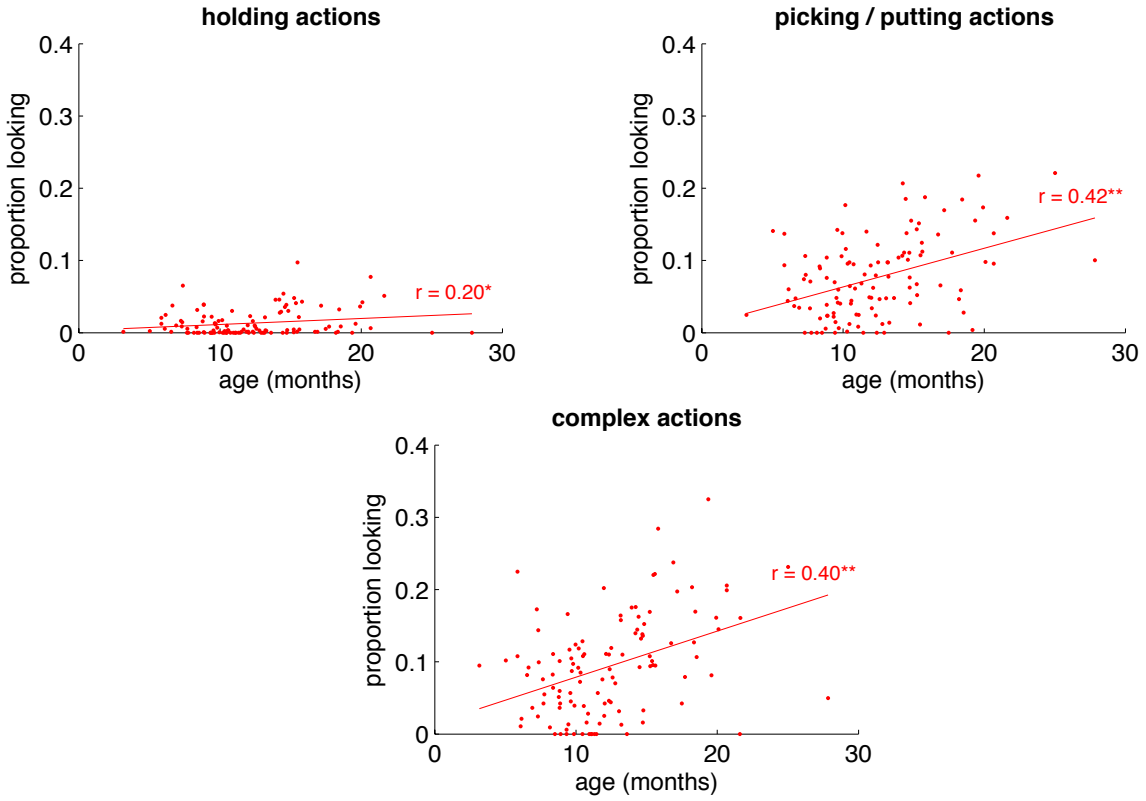

Figure 6. Each panel shows the proportion looking at hands plotted by participants' ages within a subset of the clips in the Whole Person and Multiple People conditions. Plotting conventions are as in Figure 4.

expressing emotional expressions (Figure 5). When the stimuli showed complex actions, older children looked much more at the hands of the actors performing these actions (Figure 6). These results suggest that children—especially older children—are better able to direct their social attention to the aspects of the stimulus that were most informative, given the social content.

As observed in Figure 5, all participants looked more at mouths when children in the videos were smiling or talking, and older children looked more at mouths than younger children did. These results were reflected in two mixed effects models, one for each region

of interest (because of low-level differences in area, salience, and motion, we did not compare across regions of interest). For both ROIs, we found that interaction terms did not significantly add to model fit, so we report only main effects. For the eye ROI, we found a significant negative effect of both age ($\beta = -.06$, $p = .0002$) and mouth-related actions ($\beta = -0.51$, $p = .02$). For the mouth ROI, we found significant positive effects of age ($\beta = .05$, $p = .003$) and mouth-related actions ($\beta = 1.65$, $p = .0001$). For neither model was there any significant effect of adding a coefficient related to the area of the mouth or eyes in particular videos (eyes: $\chi^2(1) = .34$, $p = .56$, mouths: $\chi^2(1) = 0.19$, $p = 0.67$), suggesting that this result was not driven by the areas of particular mouths or eyes.

As observed in Figure 6, more complex actions elicited more looking to hands, and older children looked more at hands in general. A linear mixed-effects model confirmed this impression. We first tested for an interaction between action type (holding vs. picking/putting vs. complex actions) and age but found that that a model with interaction terms did not fit better than a model with only simple main effects ($\chi^2(2) = .15$, $p = .92$), therefore we report results from the simpler model. There was a significant coefficient for age ($\beta = .037$, $p < .0017$), indicating greater hand looking as children got older. We found significant increases in hand looking for both picking/putting ($\beta = 1.09$, $p = .0016$) and complex actions ($\beta = 1.36$, $p < .0001$) compared with holding. However, these two conditions did not differ from one another (the 95% confidence intervals for these coefficients, as determined via MCMC, overlapped).

Differences in looking to hands across action types was not caused by differences in the size of hands in the videos. We calculated mean hand area, the average proportion of each frame in a clip occupied by the hand ROI, and found that it was only moderately different across the three action types (2.1%, 2.6% and 2.9%, respectively). All two-sample t-tests between hand areas for different action groups failed to reach

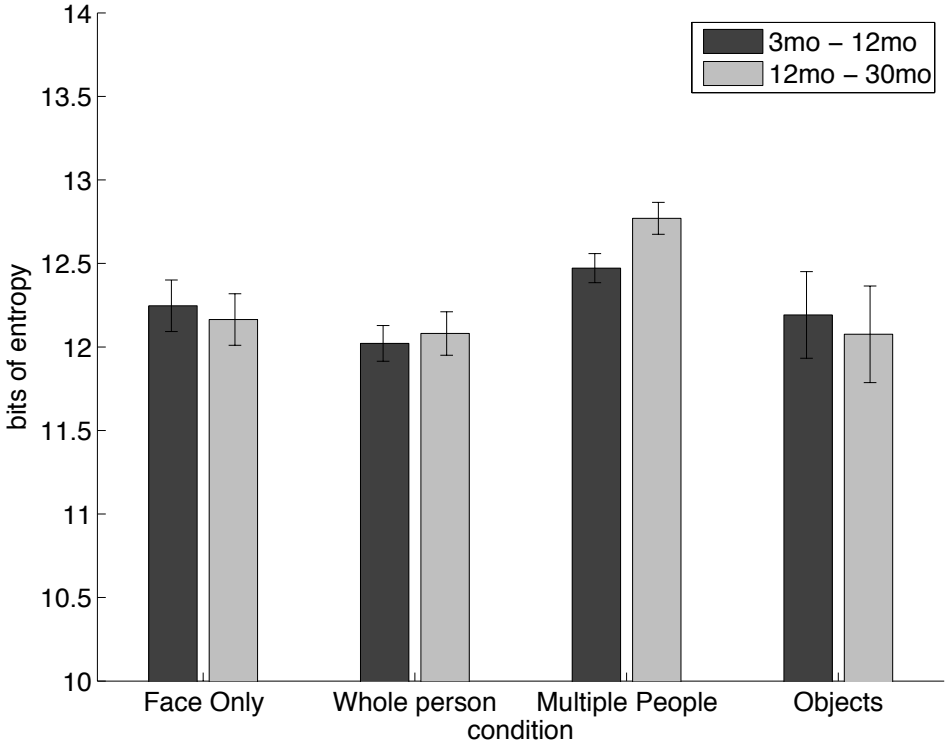

Figure 7. Average entropy of smoothed fixations across video sections for each condition. Larger entropy values indicate a broader spread of attention across participants and less predictability in fixations. Participants are split into two groups via a median split on age. Error bars show standard error of the mean.

significance, and adding hand area (ROI size) as a predictor of looking time to hands in individual clips did not significantly increase the linear model's fit ($\chi^2(1) = 1.43$, $p = .23$).

Predictability-of-fixation analyses

In addition to targeted ROI analyses, we tested whether the participants exhibited overall more constrained and predictable foci of visual attention with development. Previous results suggested that in the first 9 months of life, infants' attention goes from dispersed and unpredictable (across infants) to being relatively focused and consistent

(Frank et al., 2009). We asked whether a similar trend continued over the age range studied here by quantifying the predictability of participants' fixations. We created probability-of-fixation maps for younger and older groups of infants (median split at 11.9 months) and measured the spread of fixation within each group (see methods). For example, if all participants looked at a single face, the spread of fixation would be very limited and predictability would be very high; if each participant fixated a different location, the spread would be very broad and predictability would be low.

As seen in Figure 7, there was no consistent difference between the different groups. To quantify this impression we used a simple linear model to predict entropy across clips and age-groups. The only significant predictor was a positive coefficient on the Multiple People condition ($\beta = .42$, $p = .025$), suggesting that when there were many people, it was more difficult to predict where children would fixate. There was also a numerical trend towards lower predictability for the older children in the Multiple People condition, perhaps driven by increases in looking to hands (which would increase the number of fixation sites and hence decrease predictability). In addition, we did not see a difference in age-related predictability between the Objects control condition and the social conditions. Summing up, this analysis suggests that overall differences in gross predictability of fixation were relatively limited in the age range we examined.

General Discussion

We began our study by asking what aspects of other people draw the attention of infants and toddlers. To investigate this question, we recorded the eye-movements of a large group of infants and toddlers between 3 and 30 months as they watched engaging, live-action videos. At the highest level, our results generally confirm the findings of other studies: faces drew children's attention over other parts of the body and the surrounding physical context.

Digging slightly deeper, however, revealed developmental patterns that did not conform to expectations. The distribution of participants' fixations to faces was different both depending on their age and on what the face was doing. Younger participants looked more at the faces' eyes, while older participants looked more at mouths. This developmental difference was accompanied by an effect of content: mouth looking was overall higher when mouths were smiling or talking, even though participants could not hear what was being said. In addition, in more complex stimuli that showed adults and children performing actions, we observed a developmental shift that has not previously been reported: the older children got, the more they looked at hands, especially when the hands were involved in picking up or putting down objects or other complex actions.

Taken together, these data suggest changes in the way children view social stimuli over their first two years. The youngest infants in our sample primarily looked at faces, and within those faces, eyes. In contrast, toddlers distributed their gaze more flexibly, looking more at the sources of interesting actions and emotional expressions. This flexibility reflects greater sensitivity to social factors in the older children's looking: they were better able to allocate their social attention depending on the social context of a stimulus like a face or a hand. If the face or hand is engaged in an action that is important relative to the overall social scene then older children fixate it more than younger children; if it is simply present, then older children disengage from it more effectively.

While the effects we observed reflect the development of low-level attentional orienting abilities, they are not driven exclusively by changes in infants' non-social attention. We saw linear developmental trends across a wide range of ages, not simply in the period during the first year in which visual attention is changing most quickly (Amso & Johnson, 2006, 2008; Butcher, Kalverboer, & Geuze, 1999, 2000). In addition, although there were differences in perceptual salience between action types—for example, complex hand actions involve more motion than simple holding actions—if pure salience drove

looking at hands there would be no reason to predict developmental differences. Finally, when we examined the spread of children's fixations to moving objects, there were no gross developmental differences in the spread of fixation, suggesting that younger children were not simply more confused when looking at more complex scenes.

Previous work with younger infants documented developmental increases in looking at faces and people in complex scenes before nine months (Frank et al., 2009; Aslin, 2009). We did not see these changes outside of the Only Faces condition, but our study was not designed to detect these differences. The age range in which we had the most power was considerably older than those used in previous studies; our final sample contained primarily 8–16 month-olds, while the Frank et al. sample was from 3–9 months and the Aslin study tested 4- and 8-month-olds. In addition, the stimuli for the Frank et al. study consisted of conversations between animated cartoon characters. Both the social content and the intermodal regularities of the cartoons in the Frank et al. study supported looking at faces to the exclusion of all else; in contrast, looking at hands and mouths was often a more informative, context-sensitive behavior in our current stimuli. Thus, at a high level, infants in the earlier studies may have shown the same developmental pattern as the infants in our current study: improvements in the ability to orient to the most important part of a social stimulus (which is sometimes, but not always, the face).

Our results leave open the question of what age-related changes drive the differences we observed in our sample. While we do not believe low-level attentional factors provide a satisfactory explanation, there are myriad other developmental trends over the period we studied that could have large effects on social attention. One possible explanation is that older children simply have more experience with faces and hands. They may thus have stronger evidence about the correlation between particular visual phenomena and interesting outcomes (Triesch, Teuscher, Deák, & Carlson, 2006). Another possibility is that children's own experience with language drives looking at mouths, while their own

experiences with various types of manual actions drive looking at hands. While our own data do not distinguish between these accounts (or others of this type), we believe that examining correlations between children's experiences and abilities on the one hand and social attention on the other will be a fruitful area for much future research (Cicchino et al., 2010).

In addition to its relevance for work on social attention, the current study has implications for work across a wide variety of fields that has attempted to describe norms for eye-movement patterns in the viewing of social stimuli. This work has often used comparisons to a control group or a different age group as a way to establish population-level differences in looking patterns (e.g. Blais et al., 2008; Dalton et al., 2005; Haith et al., 1977; Klin et al., 2002; Merin et al., 2007). These efforts have made many valuable contributions to our understanding of social attention. Nonetheless, our results suggest caution in generalizing from any particular group and stimulus to predict that group's behavior with a new stimulus, and they challenge the assumption that looking at faces and eyes is always typical or healthy. Rather than only asking about the intrinsic social preferences of a group (for faces, eyes, mouths, hands, or other stimuli), we should also ask how well particular groups adapt to the unique demands presented by the social content of the stimulus.

We began by reviewing data on the robust abilities of young infants to recognize faces and make social inferences in restricted contexts. The current study, combined with previous work (Aslin, 2009; Frank et al., 2009), suggests that although these abilities may be present early, it takes time for children to display them in their moment-to-moment attention to complex social scenes. We hope that future research takes advantage of the combination of naturalistic stimuli and eye-tracking methods to continue probing the developmental trajectory of children's attention to others.

References

Amso, D., & Johnson, S. (2006). Learning by selection: Visual search and object perception in young infants. *Developmental psychology*, *42*(6), 1236–1245.

Amso, D., & Johnson, S. (2008). Development of visual selection in 3-to 9-month-olds: Evidence from saccades to previously ignored locations. *Infancy: the official journal of the International Society on Infant Studies*, *13*(6), 675.

Aslin, R. (2009). How infants view natural scenes gathered from a head-mounted camera. *Optometry & Vision Science*, *86*, 561.

Baayen, R. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge, UK: Cambridge University Press.

Blais, C., Jack, R., Scheepers, C., Fiset, D., & Caldara, R. (2008). Culture shapes how we look at faces. *PLoS ONE*, *3*(8).

Brandone, A., & Wellman, H. (2009). You can't always get what you want: Infants understand failed goal-directed actions. *Psychological Science*, *20*, 85–91.

Butcher, P., Kalverboer, A., & Geuze, R. (1999). Inhibition of return in very young infants: A longitudinal study. *Infant Behavior and Development*, *22*, 303–319.

Butcher, P., Kalverboer, A., & Geuze, R. (2000). Infants' shifts of gaze from a central to a peripheral stimulus: A longitudinal study of development between 6 and 26 weeks. *Infant Behavior and Development*, *23*, 3–21.

Cicchino, J., Aslin, R., & Rakison, D. (2010). Correspondences between what infants see and know about causal and self-propelled motion. *Cognition*.

Csibra, G., Gergely, G., Bıro, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of 'pure reason'in infancy. *Cognition*, *72*, 237–267.

Dalton, K., Nacewicz, B., Johnstone, T., Schaefer, H., Gernsbacher, M., Goldsmith, H., et al. (2005). Gaze fixation and the neural circuitry of face processing in autism.

Nature Neuroscience, *8*, 519–526.

Durand, F., & Dorsey, J. (2002). Fast bilateral filtering for the display of high-dynamic-range images. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 257–266).

Falck-Ytter, T., Gredebäck, G., & Von Hofsten, C. (2006). Infants predict other people's action goals. *Nature neuroscience*, *9*(7), 878–879.

Farroni, T., Johnson, M., Menon, E., Zulian, L., Faraguna, D., & Csibra, G. (2005). Newborns' preference for face-relevant stimuli: Effects of contrast polarity. *Proceedings of the National Academy of Sciences of the United States of America*, *102*, 17245.

Frank, M., Vul, E., & Johnson, S. (2009). Development of infants' attention to faces during the first year. *Cognition*, *110*, 160–170.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Gergely, G., & Csibra, G. (2003). Teleological reasoning in infancy: the naive theory of rational action. *Trends in Cognitive Sciences*, *7*, 287–292.

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

Gredeb ack, G., Fikke, L., & Melinder, A. (in press). The development of joint visual attention: A longitudinal study of gaze following during interactions with mothers and strangers. *Developmental Science*.

Gredebäck, G., Johnson, S. P., & Von Hofsten, C. (2010). Eye tracking in infancy research. *Developmental Neuropsychology*, *35*(1), 1–19.

Gredebäck, G., Stasiewicz, D., Falck-Ytter, T., Rosander, K., & Hofsten, C. von. (2009). Action type and goal type modulate goal-directed gaze shifts in 14-month-old infants. *Developmental psychology*, *45*(4), 1190.

Gredebäck, G., Theuring, C., Hauf, P., & Kenward, B. (2008). The microstructure of infants' gaze as they view adult shifts in overt attention. *Infancy*, *13*(5), 533–543.

Haith, M., Bergman, T., & Moore, M. (1977). Eye contact and face scanning in early infancy. *Science*, *198*, 853–855.

Hamlin, J., Hallinan, E., & Woodward, A. (2008). Do as i do: 7-month-old infants selectively reproduce others' goals. *Developmental Science*, *11*, 487–494.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85.

Hofsten, C., Dahlström, E., & Fredriksson, Y. (2005). 12-month-old infants' perception of attention direction in static video images. *Infancy*, *8*(3), 217–231.

Holland, P., & Welsch, R. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics-Theory and Methods*, *6*, 813–827.

Jack, R., Blais, C., Scheepers, C., Schyns, P., & Caldara, R. (2009). Cultural confusions show that facial expressions are not universal. *Current Biology*, *19*(18), 1543–1548.

Johnson, M., Dziurawiec, S., Ellis, H., & Morton, J. (1991). Newborns' preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, *40*, 1–19.

Kelly, D., Quinn, P., Slater, A., Lee, K., Gibson, A., Smith, M., et al. (2005). Three-month-olds, but not newborns, prefer own-race faces. *Developmental Science*, *8*, F31.

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of General Psychiatry*, *59*, 809.

Klin, A., Lin, D., Gorrindo, P., Ramsay, G., & Jones, W. (2009). Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature*, *459*(7244), 257–261.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge, UK: Cambridge University Press.

Meltzoff, A. (1995). Understanding the intentions of others: Re-enactment of intended acts by 18-month-old children. *Developmental psychology*, *31*, 838–850.

Merin, N., Young, G., Ozonoff, S., & Rogers, S. (2007). Visual fixation patterns during reciprocal social interaction distinguish a subgroup of 6-month-old infants at-risk for autism from comparison infants. *Journal of Autism and Developmental Disorders*, *37*, 108–121.

Morton, J., & Johnson, M. (1991). Conspec and conlern: A two-process theory of infant face recognition. *Psychological Review*, *98*, 164–181.

Nelson, C. A. (2001). The development and neural bases of face recognition. *Infant and Child Development*, *10*, 3–18.

Pascalis, O., De Haan, M., Nelson, C., & De Schonen, S. (1998). Long-term recognition memory for faces assessed by visual paired comparison in 3-and 6-month-old infants. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*, 249–260.

Pascalis, O., Haan, M. de, & Nelson, C. (2002). Is face processing species-specific during the first year of life? *Science*, *296*, 1321.

Quinn, P., Yahr, J., Kuhn, A., Slater, A., & Pascalis, O. (2002). Representation of the gender of human faces by infants: A preference for female. *Perception*, *31*, 1109–1122.

R Development Core Team. (2005). R: A language and environment for statistical computing. *Foundation for Statistical Computing, Vienna, Austria*.

Scaife, M., & Bruner, J. (1975). The capacity for joint visual attention in the infant. *Nature*.

Senju, A., & Csibra, G. (2008). Gaze following in human infants depends on communicative signals. *Current Biology*, *18*(9), 668–671.

Simion, F., Cassia, V., Turati, C., & Valenza, E. (2001). The origins of face perception: Specific versus non-specific mechanisms. *Infant and Child Development*, *10*, 59–65.

Triesch, J., Teuscher, C., Deák, G., & Carlson, E. (2006). Gaze following: why (not) learn it? *Developmental Science*, *9*, 125.

Walt Disney Productions. (2002). *Baby Einstein.* DVD series.

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

Yoon, J., Johnson, M., & Csibra, G. (2008). Communication-induced memory biases in preverbal infants. *Proceedings of the National Academy of Sciences*, *105*, 13690.

Yoshida, H., & Smith, L. (2008). What's in view for toddlers? using a head camera to study visual experience. *Infancy*, *13*, 229–248.