Measuring the development of social attention using free-viewing

Michael C. Frank, Edward Vul, and Rebecca Saxe
Department of Brain and Cognitive Sciences, MIT
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 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 increasingly more at hands, especially when the hands were performing actions. Our results suggest that older children’s social attention is increasingly sensitive to the content of actions.
Measuring the development of social attention using free-viewing

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 (Nelson, 2001; Gergely & Csibra, 2003). 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. Our 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. Synthesizing previous work that has been carried out using very restricted displays, 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.

We begin by reviewing work relevant to aspects of social attention. First, a wealth of research has examined infants’ and young children’s perception of faces. 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). Thus, even by six months of age,
infants have a highly sophisticated representation of faces.

Another prominent line of research has examined the perception of goal-directed actions. Results from this work suggest that as early as five months, infants encode the goal of a reach (Woodward, 1998), and that 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, Biro, 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, young children have a sophisticated understanding of others’ actions and can make inferences about the persistent mental states underlying them.

At least two other bodies of work bear on relevant aspects of social attention. First, a small literature has examined the question of visual search for faces in adults. The primary finding of this work is that search for faces among other stimuli is surprisingly efficient (Hershler & Hochstein, 2005; Lewis & Edmonds, 2003, 2005), although this pop-out effect may be due to the low-level features of faces like spatial frequency rather than some mechanism specifically tailored to face perception (VanRullen, 2006). Second, a wide variety of studies with infants and adults have investigated group-level differences in attention to faces. Cross-cultural research has found that East Asian participants looked more at noses during a face memory task, while European participants tend to fixate eyes and mouths (Blais, Jack, Scheepers, Fiset, & Caldara, 2008). Studies of infants and adults either at risk for or diagnosed with Autism Spectrum Disorders have also found that differential fixation of mouths over eyes in static and moving faces may be associated with autism (e.g. Dalton et al., 2005; Klin, Jones, Schultz, Volkmar, & Cohen, 2002; Merin, Young, Ozonoff, & Rogers, 2007).

The development of social attention is thus a central issue for researchers working in
Development of social attention

a wide variety of fields. Nevertheless, perhaps due to methodological issues, relatively little work has examined what infants and children choose to look at in natural environments or naturalistic displays. A small group of recent studies has begun to address this issue. In one study, 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.

In another study, reported by Aslin (2009), researchers gathered naturalistic recordings of one infants’ visual experience with a head-mounted camera and then recorded eye-tracking data while showing these videos to different groups of participants: 4- and 8-month-old infants and adults. They coded which regions of the videos were looked at during a variety of activities (e.g. shopping, play with blocks at home) and found age-related differences in fixation. For the most part, neither group looked at hands more than a small fraction of the time, except during blocks play. Fixations were similar across age groups, although in several contexts adults looked significantly more at people than infants did.

A final study addressed a similar question about viewing in naturalistic displays via a different approach. Frank, Vul, and Johnson (2009) gathered eye-tracking data from 3-, 6-, and 9-month-old infants as well as a control population of adults as they watched clips from _A Charlie Brown Christmas_, an engaging animated cartoon filled with social interactions between people. They 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 in looking at people observed by Aslin, 2009). In further analyses, Frank et al. found that the gaze of
younger infants was more diffuse than that of older infants and adults, and that the youngest group’s fixations were best explained by a model of motion and contrast in the movies, not by the locations of faces in the stimulus.

Why did older children look more at faces and people in both the Frank et al. and Aslin studies? First, older infants could simply have a stronger preference to look at faces due to their greater knowledge of faces’ social relevance (e.g. Triesch, Teuscher, Deák, & Carlson, 2006). Second, older infants could have been more sensitive to the intermodal regularities—temporal coordination between movement and sound—created by the presence of speech in the stimuli of the Frank et al. study (Bahrick & Lickliter, 2000; Bahrick, Lickliter, & Flom, 2004) (although there were no such regularities in the Aslin study, where stimuli were presented without audio). Third, older infants could have had the same underlying preference to look at faces or intermodally coordinated stimuli but could have had a greater ability to orient to interesting stimuli due to the development of their attentional orienting abilities.

Several factors in the design of the two studies prevented the data from distinguishing between these explanations. First, in the Frank et al. study, the stimuli were largely oriented around conversations between characters in the film; hence nearly all of the social stimuli showed intermodal regularities due to speech. Second, because the films in the Frank et al. study were cartoons rather than live action, faces were exaggerated in size and detail relative to hands and other body parts, face looking could not be compared accurately to looking to other social stimuli. Third, in both studies the range of ages tested was relatively small and was concentrated during a period in which children are undergoing large changes in their visual attention and orienting abilities (Amso & Johnson, 2005, 2006, 2008; Butcher, Kalverboer, & Geuze, 1999, 2000; Butcher, Kalverboer, Geuze, & Stremmelaar, 2002).

Our current study was designed to go beyond these previous studies and address
questions about social attention directly. As in the Frank et al. (2009) study, we set out to investigate free-viewing behavior in infants and young children, but we modified our design in several respects. First, we chose a stimulus set that did not display the same intermodal regularities as the Charlie Brown cartoon: a set of live-action movies (adapted from the Baby Einstein series) of children playing accompanied by uncoordinated classical music. Second, we systematically varied the amount of detail and complexity in our stimuli, breaking stimulus videos into four conditions: (a) children’s faces on a white background; (b) children playing with objects on a white background; (c) multiple children playing, often with adults, in a real-world setting; and (d) simple objects moving on a white or black background. This manipulation allowed us to track children’s looking to both faces and hands across a wider range of contexts. Third, we designed our recruitment procedure to include a sample of participants that was skewed towards older infants and toddlers, for whom pure oculo-motor and attentional development was more advanced.

We used this design to build upon our previous work and target three primary questions of interest:

1. How is attention allocated between faces, hands, and other targets for these age groups, and does this allocation change across development even in the absence of intermodal regularities?

2. Is the distribution of attention to socially-relevant stimuli (e.g. faces and hands) modulated by the kind of actions that are being performed?

3. Are there differences in the spread of fixation (the breath of the areas across which a group of participants fixate) between infants and toddlers even in live-action stimuli, and is the spread of fixation for these groups affected by the visual complexity of the stimuli?
Methods

Participants

Our recruiting followed an opportunistic design. Two hundred and thirty six children between the ages of zero 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 190 (80.5%) who had calibrations that could be verified or adjusted offline (see analysis section for details). Of those 190, 119 (62.6%) contributed eye-tracking data for more than 20% (48s) of the 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 119 children who fulfilled our criteria for calibration and contribution of data constituted our final sample (mean age = 12.4 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 targeted 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 4 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 torsos), 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 video 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 PlaySpace. 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. The total duration of the experiment was approximately 4m 30s.

Analysis and Results

For clarity, we divide the section reporting our analyses and results into subsections dealing with preprocessing and calibration, region-of-interest (ROI) analyses, analyses of ROI by movie content, and analyses of the spread of infants’ fixations. Unless otherwise specified, analyses were conducted using custom software written in Matlab.

Preprocessing and calibration

We first exported data from Tobii Studio and converted them into a common format. 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.

The first challenge in our analysis was verifying the calibration 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 across ages. This concern motivated the inclusion 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 (Figure 3, blue). 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). We then examined each infant’s adjusted calibration by hand 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 procedure ensured a high degree of accuracy in the calibrations of those participants included in the study.

Region-of-interest analyses

We next turn to our primary research question in this study: where children of different ages directed their gaze. The approach we took in our first analysis was a region-of-interest (ROI) approach. For each video, we used custom software to hand-code the bounding rectangle around stimuli of interest. 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 Face Only 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 each child we extracted percent dwell-time within the coded ROIs for each
Development of social attention  12

condition. Results are plotted in Figure 4. To avoid problems with sparse, noisy measurements of individual children, we excluded children from a particular condition if they did not contribute data from at least 18s (30%) of the total 1m video. When two regions of interest overlapped over a particular fixation, that fixation was counted as belonging to both regions. Supplementary videos S1–S3 show ROIs and fixations for one movie from each of the three conditions containing social stimuli (available at http://www.stanford.edu/~mcfrank/freeview).

In the Face Only condition, we found that nearly all fixation time was spent looking at faces (95.1%). 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. Nevertheless, there was still a small but significant increase in looking at faces across development ($r = .29$, $p = .004$). Turning to the eye and mouth ROIs, we saw approximately equivalent looking to each (26.2% to eyes and 19.1% to mouths). But we also observed a crossover such that younger children spent more time looking at eyes ($r = -.32$, $p = .001$) and older children spent more time looking at mouths ($r = .34$, $p = .0006$).

In the Whole Person and Face Only conditions, we saw less looking at faces (51.6% and 46.7%, respectively), though faces also made up correspondingly less of the overall area of the movie (7.2% and 10.3%). In contrast to the Face Only condition, in both conditions with more complex actions, we saw an overall decrease in looking to faces with age ($r = -.23$, $p = .02$ and $r = -.25$, $p = .02$). In addition we saw a highly consistent increase in looking to hands in both conditions ($r = .37$, $p = .0002$ and $r = .59$, $p < .0001$).

This analysis showed that younger children looked at faces (and at eyes in those faces). Older children, in contrast, looked longer at hands than younger children did and looked more at mouths than eyes.
Content analyses

In our next set of analyses we followed up on the individual ROI analyses by examining whether the content of video clips had influenced our participants’ looking behavior.

Looking at eyes and mouths. We first examined clips in the Face Only condition. We divided the video clips in this condition into two action-groups: those which included mouth movements from talking or other related vocalization (without sound) or some facial expression involving mouth movement like smiling (N=11) and those which did not (N=7). We then split the ROI data from the previous analysis on the basis of this division, taking the average looking at a particular ROI for a particular clip. Figure 5 shows these data aggregated across clips.

We next created a linear mixed-effects model (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. We chose linear mixed-effects models because of the crossed design of our data (with multiple observations for each participant and for each video clip). To control for these, we included intercept terms (“random effects”) of both participant and video clip in our models. Because average dwell times were distributed in a roughly exponential pattern, we used a logit transform to create a dependent measure that was normally distributed and hence appropriate for a linear model (we also standardized the units of dwell-time, using a z-score to increase the interpretability of coefficients). All p-values and confidence intervals reported in this and other mixed-model analyses were derived from posterior simulation using the languageR package (Baayen, 2008). The measure \( p_{MCMC} \) represents the number of samples from the model’s posterior probability distribution for which the \( \beta \) weight was in the opposite direction. This number can be interpreted as the
probability of an error in the direction of a particular effect.

We found significant main effects of looking at mouths \( (\beta = -2.93, p_{MCMC} < .0001) \) and age \( (\beta = -0.056, p_{MCMC} = .0003) \) but no significant effect of smiling/talking \( (\beta = -0.43, p_{MCMC} = .19) \). We additionally found a significant interaction of smiling/talking and looking at mouths \( (\beta = 2.19, p_{MCMC} < .0001) \). Smiling or talking increased average looking at mouths from an average of 7% to an average of 28% across ages. There was also a significant interaction of looking at mouths with age \( (\beta = .12, p_{MCMC} < .0001) \). In the smiling/talking clips, children under 12 months looked at mouths an average of 24% while those older than 18 months looked 33%. Other two- and three-way interactions did not reach significance \( (p \text{ values } > .5) \).

Summarizing the results of this analysis: all participants looked more at mouths when children in the videos were smiling or talking, and older children looked more at mouths than at eyes overall.

**Looking at hands.** For clips in the Whole Person and Multiple People conditions, we followed up on the increase in looking at hands that we observed in the ROI analysis. We divided the videos in this condition into three categories on the basis of how the hands were used: those in which the children in the videos used their hands only for holding or supporting actions \( (N=7) \); those in which hands were used for picking up an object, putting down an object, or otherwise changing its position \( (N=11) \); 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) \( (N=13) \). We again split the dwell-time data for each child on the basis of this division. The results of this split aggregated over clips are shown in Figure 6.

We again fit a linear mixed-effects model to logit-transformed and scaled dwell-times for each participant in each clip. Predictors were action (holding vs. picking/putting vs. complex actions) and age. We performed ANOVA for model comparison and found that a model with interaction terms did not fit better than a model with only simple main effects.
Development of social attention 15

\( \chi^2(2) = .16, p = .92 \), therefore we report results from the simpler model. We used holding actions as the baseline and found significant coefficients for both picking/putting actions \( (\beta = 1.09, p_{MCMC} = .0016) \) and for complex actions \( (\beta = 1.36, p_{MCMC} < .0001) \), but the 95% confidence intervals for these coefficients (as determined via MCMC) overlapped, indicating that these conditions did not differ significantly. Mean hand looking across age increased from 2% for holding actions to 8% for picking/putting and to 11% for complex actions. There was also a significant coefficient for age \( (\beta = .037, p_{MCMC} < .0017) \): for example, for infants below 12 months, looking at hands for complex actions was 8%, while for children above 18 months, it was 16%.

In order to ascertain that the differences we observed did not stem from the different size of hands across the three different action conditions, we constructed another model with the average area of hands across each clip as a further factor. The addition of this factor did not significantly increase model fit \( \chi^2(1) = 1.43, p = .23 \), but the deletion of the action categories even from a model with area included resulted in significantly reduced model fit \( \chi^2(2) = 14.30, p = 0.0008 \). In addition, mean hand sizes across the three action types were only moderately different (on average, 2.1%, 2.6% and 2.9% of frame for holding, picking/putting, and complex actions, respectively). No pairwise comparison of mean areas for different action groups (via t-test) reached significance.

Summarizing this analysis in turn, we found that older participants looked more at hands in the videos, and all participants looked more at hands when those hands were involved in picking up, putting down, or other complex actions as opposed to simply holding objects.

*Spread-of-fixation analyses*

We were interested in whether there were broad changes in how participants allocated their attention across different stimuli. In order to quantify this intuition, we...
made use of an analytic technique from Frank et al. (2009): we created probability-of-fixation maps for different groups of infants and compared their spread via entropy, an information-theoretic measure of uncertainty. This analysis measures what we call the spread of fixation: how widely the targets of fixation for different participants vary within a particular group. For example, if all participants looked at a single face, the spread of fixation would be very limited; if each participant fixated a different location, the spread would be very broad.

We implemented the analysis as follows. We first made a median split of participants by age (median = 11.9 months). We chose this split because of the distribution of ages in our dataset; since the majority of the children in our dataset were between eight and 16 months we were not able to create a larger range of age groups with equal numbers of participants for a constant change in age. In contrast, a median split ensured that the number of participants in each group was equal and gave us an opportunity to detect gross differences in spread of fixations.

We next created fixation probability maps for each of the two groups of participants. These maps were created by collecting each participants’ fixations for each frame of the stimulus and then convolving these fixations with a Gaussian kernel. 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.

We quantified the spread 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. 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 (plotted in Figure 7).

Because the number of clips in each condition was unbalanced and because we had
entropy measurements for each clip for both groups, we again used a linear mixed-effects
model to quantify the effects of age group (3 - 12mo vs. 12 - 30mo) and stimulus condition
(Faces pure, Faces medium, Faces plus, and Objects) on entropy. In order to take into
account the fact that both groups saw the same clips, we included an effect of movie clip
in the model as a random effect. We found that there was a significant effect of group on
entropy: older children had lower entropy ($\beta = -0.153, p_{MCMC} = 0.006$). In addition, there
was a relatively larger, significant effect of the Multiple People condition on entropy
($\beta = 0.809, p_{MCMC} = 0.006$) relative to entropy in the Face Only condition but coefficients
for Whole Person and Objects conditions did not reach significance ($\beta = -0.349,
p_{MCMC} = 0.19$ and $\beta = 0.171, p_{MCMC} = 0.56$, respectively). No interaction terms reached
significance (all $p$ values $> 0.39$). All coefficients can be interpreted straightforwardly in
terms of average increase or decrease in bits of entropy for a particular condition.

We found small but systematic differences in the spread of fixations across ages, but
between-condition differences were considerably larger than age-related differences.
Fixations were more broadly distributed in the Multiple People condition (which included
noisy backgrounds and multiple actors) than in any of the other conditions we examined.
Discussion

In this study 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. We were interested in how participants directed their attention in these displays and whether the distribution of attention changed across development. We found systematic changes with development in both the targets of participants’ attention and the overall distribution of attention, even in the absence of intermodal regularities.

Unsurprisingly, when large faces were presented on a white background, all participants looked at them nearly all of the time that they spent looking at the screen. However, the distribution of participants’ fixations 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 they were saying).

In the more complex stimuli in the Whole Person and Multiple People conditions, we observed a related pattern. All participants looked more at faces than at hands, but this preference was more pronounced for younger infants. For the older children we observed a substantial amount of looking at hands, but only when the hands were involved in picking up, putting down, or other more complex actions, not simply when they were holding an object.

When we examined the spread of participants’ fixations as a whole we found a small but significant decrease in entropy for toddlers over infants, as well as a more substantial increase in entropy for both groups for the stimuli in the Multiple People condition, which included more complex, real-world scenes. There were no interactions of age group and condition, however, suggesting that the younger group was not differentially affected by the more complex scenes.
Taken together, these data suggest that the way children view social stimuli changes over the first two and a half 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 at hands or mouths when the hands were involved in an interesting movement or when the mouths were involved in speech or other facial expressions. Especially in the case of hands, toddlers allocated their attention more flexibly with respect to the content of the video clips than did younger infants.

**General Discussion**

We began by asking what aspects of other people draw the attention of infants and toddlers. At the highest level, our results generally confirm the findings of other studies: faces drew children’s attention over other parts of the body. Digging slightly deeper, however, revealed a developmental pattern that did not conform to expectations. Toddlers looked less at faces and at eyes within faces than did infants. Instead, they looked at what was interesting: if children on screen were doing something with their hands or making a face, toddlers looked at their hands or their mouth. This resulted in a surprising amount of looking at mouths and faces in our oldest participants.

How should the results of the current study be integrated with previous results? In the current study, as in Frank et al. (2009), we found decreases in the spread of fixations, consistent with a developmental trend towards consistency in the foci of attention, even in complex scenes. On the other hand, we did not observe the same pattern of developmental increases in looking to faces that we saw in the 3–9 month-olds in our earlier work (which we have replicated in other ongoing work, Frank, Vul, Saxe, & Johnson, 2010). There are a number of reasons why we may not have observed this trend. First, in the final sample for our current study, there were only four children younger than six months, thus any changes that were manifest only in the youngest infants might not have been detected
here. Nevertheless, we saw no hints of such a trend, even in the Multiple People condition.

A second explanation why we might not have observed increases in looking to faces in the current study is that specific aspects of our stimulus that were not reproduced in this study might have led to the developmental changes that we observed in our earlier work. One possible change is the shift from animated characters to live-action movies. For example, it might be the case that we observed a protracted developmental course in the development of looking at faces due to the difficulty of finding the faces in animated displays. Although we know of no data that speak directly to the differences in face detection in animated versus live-action displays, previous results would suggest the opposite effect: crowded displays of higher complexity should hinder search performance (e.g. Wolfe, 1998) and schematic face displays have been at least as successful in eliciting looking preferences as real faces (Farroni et al., 2005; Johnson et al., 1991).

One final alteration in design (and the one which we believe is most likely to explain the differences between the results of the previous study and the data reported here) is the almost complete lack of speech. Our videos in the current study contained children (and sometimes adults) interacting with objects alone or together; in contrast, the Charlie Brown movie used in the earlier study was largely composed of conversations between characters. While looking at faces is more informative with respect to the content of the Charlie Brown stimulus, looking at hands is often more informative in our current materials. In other words, the older infants in our previous study could be construed as engaging in the same type of content-sensitive information selection as the toddlers in the current study.

How does this kind of information selection interact with the use of low-level regularities in stimuli? In the Charlie Brown cartoon, intermodal regularities between the moving mouths of characters’ faces and the coordinated speech of the soundtrack provided an important low-level support for infants’ ability to find and fixate their faces. In our
current study as well, participants fixated mouths and hands more when they engaged in actions, but these actions involved movement. Mouths involved in facial expressions or speech are by definition moving, and hands that are engaged in picking, putting, or complex actions are also likely to be moving more than hands which are simply holding something. Nevertheless, it is important to note first that hands were far from the only salient, moving part of the scene (especially in the Multiple People condition), and second, that if pure salience drove looking at hands there would be no reason to predict developmental differences. Thus, while lower-level perceptual regularities certainly support our results, they do not explain the pattern of data we observed.

Previous work across a wide variety of fields has attempted to describe norms for eye-movement patterns and their development in the viewing of social stimuli, often via comparison with looking patterns of a control group or a group of a different age (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 here suggest caution in generalizing from any particular group and stimulus to predict that group’s behavior with a new stimulus. 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 content of a particular stimulus.
References


Author Note

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, 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 Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Ave., Room 46-3037D, Cambridge, MA 02139, tel: (617) 452-2474, email: mcfrank@mit.edu.
Footnotes

1Only three clips appeared to include mouth movements from vocalizations (children saying “boo,” “bye,” and yelling, respectively) so we included these with other mouth-related facial expressions rather than analyzing them separately.
Figure Captions

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.

Figure 2. Three representative frames from the first video of each of the four conditions in our study.

Figure 3. An example of a single child’s eye-track on the four-point offline calibration stimulus we used. Blue dots represent individual markers of original point-of-gaze, while red dots indicate point-of-gaze after adjustment. Black crosses represent the center position of the calibration object.

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 (* = $p < .05$, ** = $p < .01$).

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.

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.

Figure 7. Average entropy of smoothed fixations across video sections for each condition.
Participants are split into two groups via a median split on age. Error bars show standard error of the mean.
The plots show the percentage looking at different body parts (eyes, faces, mouths, and hands) as a function of age (months) for three conditions: Face Only, Whole Person, and Multiple People.

- **Face Only**:
  - Faces: $r = 0.29^{**}$
  - Eyes: $r = 0.34^{**}$
  - Mouths: $r = 0.33^{**}$

- **Whole Person**:
  - Faces: $r = 0.23^{*}$
  - Hands: $r = 0.37^{**}$

- **Multiple People**:
  - Faces: $r = 0.25^{*}$
  - Hands: $r = 0.59^{**}$
no mouth expression

proportion looking

mouth

eyes

r = 0.21*

r = 0.20*

r = 0.27**

mouth

talking/smiling

proportion looking

eyes

mouth

r = 0.23*

r = 0.27**

age (months)

talking/smiling

age (months)
The diagrams illustrate the relationship between age (in months) and the proportion of looking holding and picking/putting actions, along with complex actions.

- **Holding actions**: The proportion of looking actions ($r = 0.20^*$) shows a weak positive correlation.
- **Picking/putting actions**: The proportion of picking/putting actions ($r = 0.42^{**}$) demonstrates a strong positive correlation.
- **Complex actions**: The proportion of complex actions ($r = 0.40^{**}$) also exhibits a strong positive correlation.
Face Only  Whole person  Multiple People  Objects

condition
bits of entropy

3mo - 12mo  12mo - 30mo

Face Only  Whole person condition  Multiple People  Objects

bits of entropy