The development of gender categories in infancy and childhood

Lisa C. Newell\(^1\) and Mark S. Strauss\(^2\)

Florida International University\(^1\)

University of Pittsburgh\(^2\)
Abstract

The development of gender categorization and the effect of gender-typicality (i.e., masculinity/femininity) on categorization of dynamic faces were investigated from infancy through childhood. Studies 1 and 2 tested 5-, 6-, 8-, and 11-month-old infants in a habituation task with either typical or atypical faces without cultural cues (e.g., hair or clothing). The results demonstrated that by 6 months, infants categorize the typical faces. However, even 11-month-olds were unable to categorize the atypical faces. Given the limited ability of infants to categorize atypical faces, Study 3 focused on the abilities of preschoolers. The results indicated that while 3- and 4-year-old children were able to categorize all faces above chance, they were more accurate with typical faces than atypical faces and more accurate with faces which included hair cues than those faces which had the hair occluded. This is the first set of studies to investigate the development of a single skill from infancy through childhood and indicated that while the ability to categorize facial gender begins as early as 6 months of age, it takes until late childhood to reach a level of expertise that is comparable to adults.
The infant’s world is composed of people, objects and events. What aspects of the environment are most salient to the infant? Faces have been shown to be highly salient to young infants and are an important source of information. Faces are seen more often than other stimuli and convey information about aspects of people such as emotion, intention, and identity, necessary for interacting in the social world. Newborns are able to recognize the mother’s face within hours of birth (Bushnell, Sai, & Mullin, 1989; Goren, Sarty, & Wu, 1975); they are attracted to faces from hours after birth (Goren et al., 1975; Johnson & Morton, 1991); and they track a face further than other stimuli (Goren et al., 1975). Faces are recognized better than other stimuli, even in infancy (e.g. Tanaka & Farah, 1993; Tanaka & Sengco, 1997; Thompson & Massaro, 1989; Ward, 1989); yet face recognition is an exceptionally difficult perceptual task. Faces are very complex stimuli in which minor variations have important implications for identity, emotional expression, and even gender. Despite the complexity of the task, adults demonstrate remarkable skills for perceiving and recognizing faces (e.g., Bahrick, Bahrick, & Wittlinger, 1975; Levin & Beale, 2000; O’Toole, Deffenbacher, Valentine, McKee, Huff, & Abdi, 1998). For instance, adults are able to recognize faces despite wide variations in pose, lighting, and expression (Newell, Chiroro, & Valentine, 1999). Also, individuals are able to recognize a face within ½ second, despite the fact that people have thousands of faces stored in memory (Carey, 1996).

There is a substantial amount of research, both behavioral (e.g. Tanaka & Farah, 1993; Tanaka & Sengco, 1997; Thompson & Massaro, 1989; Ward, 1989) and neuropsychological (Farah, Levinson, & Klein, 1995; Kanwisher, McDermott, & Chun, 1997; Kanwisher, Stanley, & Harris, 1999; Kanwisher, Tong, & Nakayama, 1998), suggesting that the way individuals process faces is qualitatively different from other object categories. Others suggest, however, that faces
are not processed differently than objects, but that all adults are experts in their knowledge of faces (e.g., Gauthier, Behrmann, & Tarr, 1999; Gauthier, Williams, Tarr, & Tanaka, 1998). Adults have had extensive, and meaningful, exposure to faces since birth. Gauthier and others (e.g., Diamond & Carey, 1986; Gauthier, Skudlarski, Gore, & Anderson, 2000) have conducted experiments with individuals who are experts with a particular category of objects (e.g., birds, dogs, cars), the results of which indicate that experts display similar behavioral (e.g., configural processing) and neurological (e.g., activation of the Face Fusiform Area of the Fusiform Gyrus) effects with their category of expertise as typical adults do with faces. Thus, the research with experts suggests that it is expertise or experience, not necessarily faces, which may be the critical factor in the experimental effects typically found in face research.

The hypothesis that faces are an area of perceptual expertise necessitates consideration of the role of experience and learning in face perception. Experience has been shown to be a very important factor in face perception and recognition. Previous research on face recognition in adults has identified several important findings concerning the way adults store faces in memory which seems to be a result of individuals’ experience with faces. A general framework, proposed by Valentine (1991), provides a unifying theory of face recognition that helps to explain various aspects of face recognition.

Research has shown that adults remember distinctive faces more accurately than typical faces and that they are faster at identifying typical faces than distinctive faces in a face classification task (e.g., Bartlett, Hurry, & Thorley, 1984; Going & Read, 1974; Light, Kayra-Stuart, & Hollander, 1979). In addition, adults have been shown to recognize faces of their own race better than faces from another race (Goldstein & Chance, 1980). Finally, adults have been shown to have more difficulty recognizing inverted faces than upright faces (Valentine, 1988).
Valentine (1991) proposed an experience-based face-space framework which pulls together many of these findings in the face recognition literature under a common theoretical structure. In this framework, faces are stored according to the values of their features. The framework is proposed to be an \( n \)-dimensional space representing all possible features of a face including single features (e.g., the nose), configurations (e.g., eye separation), and external features (e.g., hairline). The dimensions of the face-space framework depend upon an individual’s experience with faces, although it is an inherent assumption of Valentine’s theory that all of the distributions of features are normally distributed. Therefore, the center of this framework represents the central tendency of all the features. As an individual gains experience with faces, these faces are represented in the face-space framework according to the values of their features. With more experience, the distributions become more refined and the central tendencies become more accurate.

More typical features (and therefore more typical faces) lie closer to the center of this framework. Less typical, or distinctive, features or faces fall along the outer edges of the framework (Valentine, 1991). Since features (and therefore faces) are assumed to be normally distributed, there should be a dense cluster of typical faces in the center of the framework whereas the distinctive faces should fall in more sparsely populated regions of the framework, along the perimeter (See Figure 1). The development of a face-space framework depends on experience and implies knowledge of the range of feature values in the environment. Valentine’s theory necessitates that over time individuals be able to abstract both central tendencies and the range of values of features from their experience with faces. Research has suggested that even infants are able to abstract prototypes of faces (Quinn, 1987; Strauss, 1979) and that they can use
prototypic information to help recognize faces (de Haan, Johnson, Maurer, & Perrett, 2001) and in their perception of attractiveness (Langlois & Roggman, 1990).

Gender categorization tasks may provide the clearest model of the development of face expertise for several reasons. First, facial gender categories have a typicality distribution and thus allow researchers to investigate the impact of typicality in a face perception task. The gender-related typicality of a face is hypothesized to be determined by its distance in the face-space from a gender-specific prototype. Adults demonstrate a typicality effect in gender categorization, in which more typical faces are categorized more quickly than less typical faces (O’Toole et al., 1998). It is likely that gender is discriminated by comparing the multiple aspects of the face to separate prototypes for male and females faces (e.g., O’Toole, 1998). For typical exemplars, this comparison is relatively easy. In contrast, when trying to decide the gender of a boundary or “asexual” face, a careful decision must be made as to whether the face is closer to the male or female prototype. With development, the formation of prototypes may improve (for example, they may become more multidimensional), and thus the distribution of male versus female faces may become more distinct. Unfortunately, there is no research on this topic. In addition, gender categorization tasks do not tax memory. Discrimination of gender is dependent on memory abilities only to the extent that the individual must compare the target face to an internalized prototype or average face based on all previous experiences with faces. Finally, categorization of facial gender can be investigated with and without the availability of external features and/or cultural cues such as hairstyle and clothing, which allows researchers to unpack the relative influence of multiple cues.
Despite the suitability of gender categorization for investigating face perception skills, very little research has been conducted which investigates the development of gender categories. Previous research tends to focus on when infants are able to categorize gender, and is less concerned with the processes underlying categorization. Also, the previous research on infants’ gender categories do not control for external and/or cultural cues or the typicality of the faces.

Previous studies which investigated the development of gender categories indicated that infants categorize gender from static, silent pictures as early as 8 months of age. It is unclear to what extent infants rely on internal facial features or more simplistic cues such as clothing or hairstyle for discrimination. Yamaguchi (2000) specifically examined the ability to discriminate gender using only internal facial features. Six- and eight-month-old infants were habituated to an average male or female face. The infants were then tested with two faces that were mathematically equidistant to the habituation face; an average face of the opposite gender and a hypermale (or hyperfemale) face, matching the habituation face in gender. The hypergender faces were created by mathematically exaggerating the featural differences that have been found to exist between male and female faces (e.g., nose length and forehead size). Eight-month-old infants considered the change to the novel gender to be more novel than the change to the hypergender face (Yamaguchi, 2000). Therefore, the infants perceive a change in gender as categorically different than a mere featural change, evidence that gender categories may be developed by 8 months of age.

However, Leinbach and Fagot (1993) suggests that, even at 12 months of age, infants are not able to categorize gender on the basis of facial features. Nine- and twelve-month old infants, but not five- and seven-month old infants, were able to categorize gender as depicted in magazine pictures of individuals with gender-typical hair and clothing. However, 12-month olds
did not categorize gender when the hair and clothing cues were gender-neutral. The authors suggest that the 12-month-old infants were relying on the cultural cues of hair and clothing. An alternative interpretation is that these cues were a distraction from the facial features, and the infants would have categorized the pictures if they were removed, as in Yamaguchi (2000).

A few intermodal-matching studies have also looked at infants’ ability to discriminate gender. Walker-Andrews, Bahrick, Raglioni & Diaz (1991) and Lewkowicz (1996) found that infants around 5 to 6 months of age are able to match the gender of the voice to the appropriate face when presented with dynamic videos. Poulin-Dubois, Serbin, Kenyon, and Derbyshire (1994) found that 9- and 12-month-old infants matched face and voice with the female, but not the male, stimuli. However, Poulin-Dubois et al. (1994) used black and white photographs, rather than color videos. This difference suggests that infants find dynamic videos easier to discriminate. Another possibility is that the infants preferred to look at the female faces, regardless of the voice presented. Other studies have demonstrated that infants spend more time looking at female versus male faces (Brooks-Gunn & Lewis, 1981, Quinn, Yahr, Kuhn, Slater & Pascalis, 2002).

While there are several studies which address the ability of infants to categorize gender, this research has not tackled the most relevant issues identified by the adult literature. First, as just reviewed, current research has shown that the ability is developing sometime during the first year, but a more specific age is still unknown. Second, most of the studies do not control for external cues to gender such as hair or clothing. Third, the majority of studies use photographs rather than dynamic videos. Finally, none of the studies have addressed whether the typicality of the faces affects the ability to categorize gender. The importance of the three remaining issues is discussed.
Role of facial features versus other cues

One of the most basic issues concerning the categorization of gender is the basis of infants’ gender discrimination. Do infants discriminate gender using internal facial features, external cultural cues, or both? Research has suggested that adults are able to categorize gender using only the facial features (Bruce & Young, 1998; O’Toole et al., 1998). The current research, however, has not adequately addressed whether infants categorize gender when only the facial features are available. Leinbach and Fagot (1993) controlled for cultural cues, but it is possible that the faces were not salient within the context for the infants. Also, these results conflict with those found by Yamaguchi (2000), which indicated that infants categorized gender using internal facial features well before 12 months of age. Other studies have not controlled for hair or clothing cues (Walker-Andrews et al., 1991; Lewkowicz, 1996; Poulin-Dubois et al., 1994; Levy & Haaf, 1994).

Previous research has shown that infants are attending to faces and show a preference for faces over equally complex stimuli within hours after birth. Eye-tracking studies suggest that much of the infant’s attention is focused on the outer contours when viewing a face (Maurer & Salapatek, 1976). In addition, newborns recognize their mother’s face only if the hairline/hairstyle is not occluded (Pascalis, de Schonen, Morton, Dereulle, & Fabre-Grenet, 1995). On the other hand, gender categories are not perfectly correlated with hairstyle or clothing type and infants may base their categories on facial features only. Thus, the basis of infants’ gender discrimination has yet to be determined and will be addressed by the current collection of studies.
Stimulus quality

Studies that use dynamic videos (Walker-Andrews et al., 1991; Lewkowicz, 1996) suggest gender categorization at younger ages than studies which use static pictures (Leinbach & Fagot, 1993; Poulin-Dubois et al., 1994; Yamaguchi, 2000). There are several reasons why videos may be better than pictures to measure infants’ face perception abilities. Perception research has shown that features are more salient when objects are in motion. For example, motion enables infants to perceive partially occluded objects as a whole (e.g. Spelke, 1988) or to discriminate point light displays as a walking human (e.g. Bertenthal, Proffitt & Cutting, 1984). Motion may also enhance gender categorization. Motion makes features such as the height and depth of cheekbones more apparent. Arterberry et al. (2001) found that adults are better at discriminating gender in dynamic versus static stimuli and that some gender cues, such as eye/eyebrow distance, are more apparent in dynamic stimuli. Moreover, research has shown that infants prefer moving stimuli more so than their static counterparts, and are better able to detect features in moving stimuli (Slater & Butterworth, 1997).

Typicality of faces

As with other natural categories, gender categories have a typicality distribution. Some faces are considered more gender-typical than others (O’Toole et al, 1998). As stated above, adults are much quicker at judging the gender of more typical faces than less typical faces (O’Toole et al., 1998). Thus, the masculinity (or femininity) of a face may also be a factor in infants’ abilities to categorize gender. O’Toole et al. (1998) proposed that individuals store faces according to gender-specific prototypes, in two different “face spaces”. If faces are stored based on gender-specific prototypes, then the distance from the prototype is indicative of how masculine/feminine a face is. Hence, faces that are more gender-typical are closer to the
prototype and more quickly classified. It is unknown whether such typicality differences affect infants’ abilities to categorize gender. If infants are processing and storing faces according to a face-space framework, they should also display a typicality effect in gender categorization. Thus, when infants are first learning the gender categories, they should be able to discriminate only very typical exemplars due to emerging category structures. The current research has not controlled for typicality of stimuli. Poulin-Dubois et al. (1994) used only the most typical faces, but did not compare performance with the least typical faces. If there is a typicality effect, infants should be able to categorize the most typical faces at an earlier age than they can categorize the least typical faces.

Since no previous study has controlled for these discussed issues, the following set of studies were designed to investigate the role of internal facial features and typicality, using dynamic stimuli, to yield a clearer picture of the development of gender categorization.

Study 1

The purpose of Study 1 was to determine whether 5- and 8-month-old infants are able to categorize gender based on only internal facial features. Also, the study was designed to determine if the ability to categorize gender during infancy is affected by the typicality of the faces, as it is during adulthood. The design used dynamic videos of faces in which both hair and clothing cues eliminated.

Method

Participants. The participants were 40 5-month-old infants (M=152.45 days, SD=13.66) and 40 8-month-old infants (M=238.80 days, SD=14.95); each group had equal numbers of male and female infants. All infants were full-term, with no major pregnancy or birth complications. Forty-six infants were eliminated from data analyses due to fussing (n=11), experimenter or
computer program error \((n=16)\), failure to reach habituation criterion \((n=18)\) and external interference \((n=1)\). Infants’ names were obtained from a marketing company and reflected the cultural distribution of the area.

**Stimuli.** Approximately 80 videos were made of male and female individuals ranging in age from 18-30 years old. The videos were taken with a digital camcorder and downloaded into a computer. Stimuli volunteers were required to wear a common black robe to hide their clothing. To hide their hair, volunteers wore a basic black baseball cap backwards, with the hair drawn to the back. When the videos were taken, they were framed so that just the face and a minimal amount of border were included on the video. The volunteers were seated in front of a black curtain, which blended with the cap and robe. With the robe, the cap, and the background all the same color, the videos provided a very dynamic display of just a face. Hoping to elicit a natural poise from the volunteers and to make the video as realistic as possible for the infants, volunteers recited a common nursery rhyme (Hickory, Dickory, Dock) during filming. Also, volunteers were asked to pretend that they were talking to a picture of a baby, which was held beside the camcorder during filming. The videos were initially recorded with sound, but the sound was later digitally removed to eliminate voice as a possible cue to gender.

Twenty undergraduate students rated each of the 80 videos for typicality of gender on a 7-point scale, with 1 being very atypical of that gender and 7 being very typical of that gender (i.e. very masculine or very feminine). The 10 most typical female videos (to be referred to as “typical”; \(M=4.45, SD=0.42\)), the 10 most typical male videos \((M=4.73, SD=0.22)\), the 10 least typical female videos (to be referred to as “atypical”; \(M=2.45, SD=0.39\)), and the 10 least typical male videos \((M=3.48, SD=0.33)\) were selected. The least typical faces (“atypical”) were then presented to a second group of undergraduate students who were asked to identify the gender of
each face to ensure that all of the faces were easily discriminated by adults. Thus, while not as
typical of gender as the “typical” faces, the gender of the “atypical” faces could easily be
discriminated by adults. *T*-tests indicated that the ratings for the typical and atypical faces were
significantly different from each other, for both the male \[t(19)=10.03, p<.001\] and the female
\[t(19)=11.44, p<.001\] videos.

**Apparatus.** Infants were tested in an enclosed booth, seated on their caregiver’s lap. The
caregivers were instructed not to interfere with the infants’ behavior in any way. A sound
machine, playing white noise, was used to block extraneous sounds. The laboratory was
completely dark during testing. The infant was seated approximately 45-60 cm away from a 56-
cm monitor. The stimuli and inter-trial attention grabber were presented on this monitor.
Mounted above the monitor was a digital camcorder with infrared capabilities, which could not
be seen by the infant. The infrared capabilities allowed the experimenters to clearly see visual
fixations in complete darkness. The visual fixations were monitored on a television screen. An
experimenter pressed a mouse button whenever the infant was looking at the monitor. Looking
times were recorded on a computer with a program designed to run the habituation paradigm.
Testers were undergraduate students who, while trained in infant testing procedures, were
generally naïve as to the hypothesis of the study. Additionally, testers were unable to see the
stimuli during the trials so looking times were determined completely by observing the infants’
fixations.

**Procedure.** Infants were tested using an infant-controlled habituation paradigm. The
program first presented a moving 3-D tunnel-like design to attract the infant’s attention to the
computer monitor. Once the infant was looking at the monitor, a video of a face was shown.
The video stayed on until the infant looked away for one second. Once the infant looked away,
the attention grabber came back on. When the infant re-oriented, a new video was presented (e.g., a second person). The program continued like this, recording looking times, until the infant was habituated. The infant was considered to be habituated when the average of any three consecutive trials was less than half the average of the first three trials. The program required a look to be at least 0.5 seconds in order to count towards establishing or reaching habituation criterion. Once the infant was habituated, s/he was presented with eight test trials, consisting of alternating videos of male and female faces. If the infant did not reach habituation within twenty trials, the program automatically proceeded to test trials. Ultimately it was decided that these subjects would not be included in the final data analyses.

Infants were assigned to one of four habituation conditions, based on the gender (male/female) of the faces and the typicality (typical/atypical) of those faces. Therefore, they were habituated to either: 1) typical males, 2) typical females, 3) atypical males, or 4) atypical females. Infants who were habituated to typical faces were shown typical faces during test trials, and conversely, infants who were habituated to atypical faces were shown atypical faces during test trials. There were eight test trials, consisting of alternations of two male faces and two female faces, repeated twice. Eight test trials were used to try to obtain a more complete measure of infants’ novelty preferences. Ten random orderings were compiled for each habituation condition. Infants were counter-balanced across conditions.

Results

Habituation Phase. Habituation data were analyzed to determine if infants displayed any differences in looking times with respect to the age or gender of the infants or the typicality or the gender of the habituation stimuli. Three different dependent measures of habituation were analyzed. These were the average length of the first three looks during habituation, the average
length of the time spent looking per trial and the average number of trials it took to habituate. Prior to analysis, all looking times were transformed into logarithms in order to normalize the data. Each of these characteristics was examined in a 2(age) x 2(typicality) x 2(infant gender) x 2(stimulus gender) ANOVA.

Analyses of the infants’ average looking time during the first three habituation trials indicated that the 5-month-old infants had longer initial looking times ($M=2.50$ s, $SD=0.73$) than the 8-month-olds ($M=2.18$ s, $SD=0.64$), $F(1,78)=4.72$, $p<.05$. It was also found that, across the age groups, infants initially spent more time looking at the female habituation videos ($M=2.54$ s, $SD=0.76$) than the male habituation videos ($M=2.15$ s, $SD=0.58$), $F(1,78)=6.77$, $p<.05$.

Similar results were found with respect to the average looking time per habituation trial. Infants who were habituated to female videos spent more time looking at each video ($M=2.44$, $SD=0.71$) than infants who were habituated to male videos ($M=1.89$, $SD=0.57$), $F(1,78)=6.63$, $p<.05$. There were no other significant main effects or interactions with respect to average looking time per trial.

Analysis of the number of trials to habituation yielded results complimentary to the other analyses. Two significant results were found from a 2(age) x 2(typicality) x 2(infant gender) x 2(stimulus gender) ANOVA. A main effect of age indicated that the 8-month-olds required more trials to reach habituation ($M=9.95$, $SD=4.44$) than the 5-month-olds ($M=7.95$, $SD=3.55$), $F(1,78)=4.87$, $p<.05$. There was also an age x habituation stimulus gender interaction, $F(1,78)=7.60$, $p<.01$. $T$ tests indicated that the number of trials required for 5-month-old infants habituated to females to reach criterion was significantly different from both the number of trials required for 5-month-old infants habituated to males ($t=3.83$, $p<.001$) and the number of trials for 8-month-old infants habituated to females ($t=-3.73$, $p<.001$).
**Test Phase.** Log transformations were performed on the individual looking times to normalize the scores and reduce the variance. The four familiar test trials were combined and compared to the looking times for the four novel test trials. This provided scores of total looking to familiar trials and total looking to novel trials. Figure 2 shows the mean looking times and standard deviations for familiar and novel test trials for each age and typicality condition. A 2(trials; novel vs. familiar) x 2(age) x 2(typicality) Repeated-Measures ANOVA was performed on the looking time during test trials. Results indicated a main effect of novel vs. familiar trials, $F(1,78)=5.01, p<.05$, a trials x typicality interaction, $F(1,78)=4.48, p<.05$, a trials x age interaction, $F(1,78)=5.18, p<.05$, and a trials x typicality x age interaction, $F(1,78)=8.56, p<.005$. To clarify the 3-way interaction, further analyses were conducted separately on the 5- and 8-month-old groups.

A 2(trials) x 2(infant gender) x 2(stimulus gender) Repeated-Measures ANOVA for the 8-month group yielded no significant main effects of interactions, indicating that neither gender of the infant nor gender of the habituation stimuli were mediating factors in looking times. Therefore, gender of the infants and the stimuli were combined for subsequent analyses. A 2(trials) x 2(typicality) Repeated-Measures ANOVA for the 8-month group yielded a significant main effect of trials, $F(1,39)=9.26, p<.005$, indicating that the infants looked longer at the novel trials than the familiar trials. A significant trials x typicality interaction, $F(1,39)=11.56, p<.005$, was because the infants in the typical condition showed a significant preference for the novel gender, $t(39)=4.74, p<.001$, the infants in the atypical condition did not show a novelty
preference $t(39)=-0.24, p>.10)$. Thus, infants in the typical condition categorized the gender of the faces; however, infants in the atypical condition did not categorize gender.

A 2(trials) x 2(infant gender) x 2(stimulus gender) Repeated-Measures ANOVA for the 5-month group yielded no significant main effects or interactions, indicating that all infants performed similarly. A 2(trials) x 2(typicality) Repeated-Measures ANOVA for the 5-month group also yielded no significant main effects or interactions. Thus, the 5-month-old infants were not able to categorize gender with either the typical or the atypical faces.

Discussion

Study 1 revealed that both the five- and eight-month-old infants displayed a preference for female faces over male faces during the habituation phase of the procedure. The preference for female faces is a consistent finding that will be discussed in more detail in the General Discussion section. More importantly, the study indicated that, at five months, infants are unable to categorize the gender of faces when hair and clothing cues are removed, regardless of the typicality of the faces. However, by eight months, infants were able to categorize gender on the basis of internal facial features, but only for the typical faces. Thus, a typicality effect in gender categorization exists during infancy. It appears that infants first are learning to categorize the typical faces, and sometime later in development they learn to categorize the atypical faces.

Study 2

Study 2 was designed to further elucidate the development of gender categorization during infancy. The study tested 11-month-old infants in the same procedure, using only atypical faces, to determine if infants are able to categorize these atypical faces by the end of the first year. Study 2 also tested 6-month-old infants with only the typical faces to determine more precisely when the ability to categorize the typical faces develops.
Method

Participants. The participants were 22 6-month-old infants (M=190.55 days, SD=11.96), \( n=10 \) females and 21 11-month-old infants (M=336.19 days, SD=11.21), \( n=10 \) females. The infants were all full-term, with no major pregnancy or birth complications. Nineteen infants were eliminated from data analyses due to fussing (\( n=10 \)), experimenter/program error (\( n=4 \)), and failure to reach habituation criterion (\( n=5 \)). Infants’ names were obtained in the same way as the names from Study 1. None of the infants had participated in Study 1.

Stimuli. The stimuli were identical to those used in Study 1.

Apparatus. The infants were tested in the same manner as before, using the same apparatus.

Procedure. Infants were tested using the same computer program as used in Study 1 to run the infant-controlled habituation paradigm. The same orderings of videos as before were used, with the same four conditions. However, based on the results from Study 1, 6-month-old infants were only tested in the typical condition and 11-month-old infants were only tested in the atypical condition.

Results

Habituation Phase. Looking times during the habituation phase were analyzed to determine if there were any differences between gender of the tested infants or male and female stimuli prior to test trials. In order to be consistent across studies, the same three dependent measures as used in Study 1 (i.e., total number of habituation trials, the average length of the first three habituation trials, and the average amount of time spent looking per habituation trial) were analyzed. All looking times were log converted prior to analyses. Each of these characteristics
was examined in a 2(infant gender) x 2(stimulus gender) ANOVA, with separate ANOVAs for each age group. Because the 6- and 11-month-olds infants were presented with different stimuli, all analyses were performed separately for 6- and 11-month olds.

Analysis of the number of habituation trials for the 6-month-old infants in a 2(infant gender) x 2(stimulus gender) ANOVA revealed that there were no main effects or interactions for either infant gender or stimulus gender ($p>.10$). With regard to the average length of the first three habituation trials, a 2(infant gender) x 2(stimulus gender) ANOVA for the 6-month group indicated a main effect of stimulus gender, $F=5.17, p<.05$. This was a result of longer initial looking times to the female videos ($M=2.93, SD=0.59$) than the male videos ($M=2.39, SD=0.39$). Finally, analysis of the average length of looking time per trial indicated a marginal main effect of stimulus gender, $F=3.92, p=.06$. This result was also due to longer looking times for those infants who were habituated to the female videos ($M=2.56, SD=0.52$) than male videos ($M=2.11, SD=0.45$).

Analyses of the dependent measures of habituation for the 11-month-olds yielded no significant main effects or interactions. There was no effect of either infant gender or stimulus gender for the total number of habituation trials, the average looking time for the first three trials, or the average length of looking time across all habituation trials, $p>.10$.

**Test Phase.** Log transformations were again performed on the individual looking times to normalize the scores and reduce the variance. As before, the four familiar test trials were combined and compared to the looking times for the four novel test trials. This provided scores of total looking to familiar trials and total looking to novel trials. Figure 3 shows the mean looking times and standard deviations for familiar and novel test trials for each age. Again,
separate analyses were performed for each age group because they were tested with different stimuli (i.e., typical faces for 6-month-olds and atypical faces for 11-month-olds).

The 2(trials) x 2(infant gender) x 2(stimulus gender) Repeated-Measures ANOVA for the 6-month group yielded a significant main effect of trials, $F(1,21)=9.84, p<.01$ indicating that infants looked longer at the novel trials ($M=8.14, SD=1.67$) than the familiar trials ($M=7.14, SD=1.34$). No other main effects or interactions were significant, which indicates that 6-month olds are able to categorize the gender of typical faces, with neither infant gender nor stimulus gender as mediating factors.

The 2(trials) x 2(infant gender) x 2(stimulus gender) Repeated-Measures ANOVA for the 11-month group yielded no significant main effects or interactions. The lack of a main effect of trials indicates that the 11-month olds were not able to categorize gender with the atypical faces.

Discussion

The results from Study 2 reveal that the ability to categorize the gender of typical faces emerges by 6 months of age. Taken together with the results of Study 1, this indicates that the ability to categorize typical faces, as measured by these dynamic stimuli, develops during the period between 5 and 6 months of age. The ability to categorize the gender of the atypical faces, however, has not yet developed by 11 months of age. This result indicates that the ability to categorize the gender of the atypical faces is a much more difficult task that takes a year or more experience with faces to emerge. Thus, the results of Studies 1 and 2 clearly indicated that infants are able to categorize gender using only internal facial features by the end of the first year. However, they can do so only when the faces are typical exemplars of men and women.
At issue is when children are able to categorize atypical exemplars. Thus, the next study was designed to test the ability of much older three- and four-year-old children to categorize gender when the exemplars are both typical and atypical.

Study 3

Since all the stimuli used in the infant studies were categorized by adults with near perfect accuracy, preschoolers were tested to investigate when the categorization of atypical faces emerges, and how it develops. In addition, previous research has suggested that children begin to learn the cultural aspects of gender such as gender labels and gender-stereotyped behaviors (e.g., ironing or hammering) during the second year of life (Poulin-Dubois, Serbin, & Derbyshire, 1998; Poulin-Dubois, Serbin, Eichstedt, Sen, & Beissel, 2002; Serbin, Poulin-Dubois, & Eichstedt, 2002). Thus, Study 3 manipulated the availability of hair cues to assess whether preschoolers use this cultural cue as an aid in a gender categorization tasks. The stimuli were silent videos of individuals with the clothing cues occluded. Both typical and atypical faces were included in the stimulus set. Also, half of the stimuli included hair cues, while the other half had the hair occluded.

Method

Participants. The participants were 24 3-year-old children ($M=3$ years, 6 months, $SD=3.50$ months) and 24 4-year-old children ($M=4$ years, 4.66 months, $SD=3.08$ months). There were 12 female and 12 male children in each age group. All children were typically developing, with no major diseases or disorders. Twenty-six children were eliminated from data analyses for failure to complete all 40 trials. Infants’ names were obtained from a marketing company and reflected the cultural distribution of the area.
**Stimuli.** The stimuli were identical to the videos used in the infant studies.

**Apparatus.** Children were tested in a quiet room in front of a 38-cm computer monitor. The caregiver was present, but was asked not to interfere with the child’s answers in any way. The child was seated approximately 45-60cm away from the computer monitor. Videos were presented sequentially, in a random order. The video presentations were controlled by a computer program designed to present stimuli sequentially and stay on until the experimenter recorded the answer by pressing the appropriate button on the keyboard.

**Procedure.** Children were first tested in a pre-test to confirm that they understood the task and the terms being used by the experimenter. Parents were questioned before testing about the child’s use of gender terms and the experimenter used whichever words the parent indicated the child used most often. For instance, some children used the terms “man” and “woman” or “mommy” and “daddy” instead of “boy” and “girl.” Children were shown six full-body pictures from magazines and asked to identify the gender of the model in each picture. If the child incorrectly identified the gender of a picture it was repeated at the end of the pre-test. If the child incorrectly identified the gender of a picture a second time, the data from that child was excluded from the final dataset.

There were a total of 40 videos shown to each child for the main testing procedure. Children were asked to identify the gender of each video using the terms the child used everyday. Each set of videos contained equal numbers of male and female faces, typical and atypical faces, and videos with and without hair cues. Therefore gender, typicality, and hair cues were all within-subjects variables.
Results

Accuracy scores were computed based on the number of correct responses given by the child divided by the total number of completed trials. In order to determine whether preschoolers were able to reliably categorize all faces, accuracy scores from each group of stimuli (i.e., typical with hair, typical with cap, atypical with hair, atypical with cap) were entered into one-sample $t$ tests. All accuracy scores were significantly greater than chance ($50\%$), $p<.05$. Figure 4 shows the means and standard deviations for both three- and four-year-olds for each stimulus group, as well as the eight- to twelve-year-old group from Strauss, Newell, Giovannelli, Best, and Minshew (submitted).

Accuracy scores were entered into a $2$(typicality) x $2$(hair cues) x $2$(age) Repeated-Measures ANOVA. Results indicated a main effect of typicality, $F(1,46)=42.28$, $p<.001$. Overall, children were more accurate with the typical faces ($M=.84$, $SE=.02$) than the atypical faces ($M=.72$, $SE=.02$). There was also a main effect of hair cues, $F(1,46)=66.46$, $p<.001$, indicating that across both ages and for both typical and atypical faces, children were more accurate with faces which included hair cues ($M=.85$, $SE=.02$) than those which had the hair occluded ($M=.71$, $SE=.02$). Finally, there was a main effect of age, $F(1,46)=5.06$, $p<.05$ indicating that 4-year-olds were significantly more accurate ($M=.81$, $SE=.02$) than 3-year-olds ($M=.74$, $SE=.02$) across all stimulus groups. There were no other main effects or interactions.

Discussion

The results of Study 3 indicated that 3- and 4-year-old children reliably categorize gender. Therefore, important conceptual development occurred between 11-month-old infants
and 3-year-old children. However, the 3- and 4-year-old children have not yet reached adult-level categorization abilities, as all of the faces in the study are categorized by adults at 100% accuracy. It is surprising that even after 4 years of experience with faces; children are still not categorizing the gender of faces at 100% accuracy. However, 4-year-old children categorized all the stimuli more accurately than 3-year-old children, indicating that significant development is still taking place during the preschool period. Figure 4 illustrates the accuracy scores of typically developing 8-year olds in an identical task from Strauss et al. (submitted) for comparison with the 3- and 4-year-old children. By eight years of age, children are categorizing the gender of faces at or near 100% accuracy, similar to adults. Thus, gender categorization abilities have fully developed by eight years of age.

General Discussion

The current series of studies begins a systematic investigation of the development of gender categorization during infancy and early childhood. The results from Study 1 suggested that 8-, but not 5-, month-old infants categorize the gender of the most typical faces when external and/or cultural cues (including hair) are excluded from the stimuli. Study 2 followed up these results to demonstrate that the ability to categorize typical faces develops as early as 6 months. However, infants were unable to categorize atypical faces at 11 months. Study 3 indicated that preschoolers categorize gender significantly above chance levels, but still show evidence of an accuracy advantage for typical faces. In addition, Study 3 suggested that by 3- and 4-years of age, children were attending to hair cues and the presence of hair cues assisted them in gender categorization tasks.
This is the first published study to track the development of a skill from infancy through early childhood. What information can be taken from the current series of studies to enhance our knowledge of the development of gender categories specifically and face perception in general? The most remarkable aspect of this series of studies is that despite the need for different methodologies for the two age groups, there is remarkable consistency in results. Across all age groups tested, it was found that typical faces were categorized better than atypical faces. Thus, integrating these results with results from previous studies it was found that from infancy through adulthood, gender categorization is influenced by typicality. Typical faces are categorized first in development, and continue to be categorized most quickly and accurately throughout the lifespan (O’Toole et al., 1998). It is important to note that previous studies of the development of gender categories have found that gender categories develop in the first year. No other studies have asked whether gender categorization continues to improve with development. Studies investigating early childhood address children’s knowledge of verbal labels of gender and gender stereotypes. Previous research of infant gender categories has likely used mostly typical faces. If only typical faces had ever been investigated in gender categorization tasks and the limits of this skill had never been pushed, one would conclude that gender categorization develops during the first year and does not improve after infancy. However, it was demonstrated in the current studies and Strauss et al. (submitted) that while gender categories begin to emerge at six months, they continue to develop until eight to twelve years.

Why are typical faces easier to categorize? One obvious reason may be that typical faces are seen more frequently in the environment and thus are learned first and best. However, based on this theory, one would predict young infants to learn the faces most closely related to the typicality of their immediate experience, such as parents and other close relatives. If a particular
infant interacts most frequently with faces that are less typical, that child should learn to categorize less typical faces first. While further research would be required to explore this hypothesis, it seems an unlikely explanation. The most striking aspect of research on typicality distributions in categorization is the amount of agreement among people as to what represents a typical exemplar (Rosch, 1978).

Another, more likely hypothesis is that different processes may be required to categorize more versus less typical faces. As faces become less typical, they may require more configural processing and encoding of multiple features to be categorized. Previous research investigating the processes used in face perception has suggested that the efficiency of configural processing increases with development (Cashon & Cohen, 2004; Schwarzer, 2000). More typical faces may be discriminated on the basis of more featural information. For instance, a typical male face may be categorized by the size of the eyebrows or the width of the chin. However, a less typical face may require more configural processing. For example, a less typical male face may require encoding the amount of separation between the eyebrows and the length of the nose and the height of the forehead. The increased use of configural processing with development may account for the results in the current studies that less typical faces are learned later in development. These less typical faces may not be categorized until configural processing is learned.

In addition to configural processing, less typical faces may require attending to multiple features/configurations for categorization. As with configural processing, the ability to attend to multiple features simultaneously likely improves with development. This improvement in attentional strategies can also account for the evidence of a typicality advantage in gender categorization across development. Furthermore, individuals may not only need to use
configural processing strategies and attend to multiple features simultaneously, but the features/configurations that are most useful for categorization may change with each face that is encountered. For instance, one face may be categorized as a male face because it has a wide chin and a large forehead. However, another face may be categorized as a male face because it has a long, wide nose; big, bushy eyebrows; and a prominent jaw line. The ability to categorize both of these faces requires an individual to flexibly attend to different features/configurations of a face to determine which is most similar to the gender-specific prototype. Thus, improvements in face processing abilities may involve more than just a shift in emphasis from featural to configural processes but also an increase in flexibility in weighting properties depending on the circumstances.

The results from the current collection of studies in conjunction with previous research on face perception suggests that the role of typicality in all face perception tasks, the relative influence of internal versus external facial features, and the development of attentional flexibility in face processing are important issues in development face perception research. Future research will need to investigate what features and/or configurations are used to categorize or recognize faces, whether the number of features increases with development and whether the specific features and configurations encoded are different for every face.
References


Figure Captions

Figure 1. Schematic representation of Valentine’s face-space framework.

Figure 2. Looking times by age group and typicality condition for all eight test trials, Study 1.

Figure 3. Looking times by age group for all eight test trials, Study 2.

Figure 4. Accuracy scores for 3-, 4-, and 8- to 12-year-old children, for Study 4 (3- and 4-year-olds) and Strauss et al. (8- to 12-year-olds; submitted).
Figure 1.
Figure 2.
Figure 3.
Figure 4.