Audio-Visual Speech Perception in Infants and Toddlers With Down Syndrome, Fragile X Syndrome, and Williams Syndrome

Dean D'Souza[1], Hana D'Souza[1,2], Mark H. Johnson[1], Annette Karmiloff-Smith[1*]

[1] Centre for Brain and Cognitive Development, Department of Psychological Sciences, Birkbeck, University of London, London, UK.

[2] Department of Experimental Psychology, University of Oxford, Oxford, UK.

* Corresponding author

E-mail: a.karmiloff-smith@bbk.ac.uk

Abstract

Typically-developing (TD) infants can construct unified cross-modal percepts, such as a speaking face, by integrating auditory-visual (AV) information. This skill is a key building block upon which higher-level skills, such as word learning, are built. Because word learning is seriously delayed in most children with neurodevelopmental disorders, we assessed the hypothesis that this delay partly results from a deficit in integrating AV speech cues. AV speech integration has rarely been investigated in neurodevelopmental disorders, and never previously in infants. We probed for the McGurk effect, which occurs when the auditory component of one sound (/ba/) is paired with the visual component of another sound (/ga/), leading to the perception of an illusory third sound (/da/ or /tha/). We measured AV integration in 95 infants/toddlers with Down, fragile X, or Williams syndrome, whom we matched on Chronological and Mental Age to 25 TD infants. We also assessed a more basic AV perceptual ability: sensitivity to matching vs. mismatching AV speech stimuli. Infants with Williams syndrome failed to demonstrate a McGurk effect, indicating poor AV speech integration. Moreover, while the TD children discriminated between matching and mismatching AV stimuli, none of the other groups did, hinting at a basic deficit or delay in AV speech processing, which is likely to constrain subsequent language development.

212 words

Key words: Down syndrome, fragile X syndrome, Williams syndrome, audio-visual speech integration, language acquisition, the McGurk effect

Highlights

- Audio-visual (AV) speech integration is important for language development
- AV speech integration was assessed in children with neurodevelopmental disorders

- AV integration is impaired in infants with Williams syndrome
- General AV speech processing skills may be impaired in Down and fragile X syndromes

Audio-Visual Speech Perception in Infants and Toddlers With Down Syndrome, Fragile X Syndrome, and Williams Syndrome

Children acquire their native language quickly and with relative ease. How they achieve this remarkable feat is not fully understood. One potential and oft neglected mechanism involves the visual speech cue. Visual speech has often been regarded as a redundant signal in verbal communication (Weikum et al., 2007). But although it has been shown that children can discriminate languages both auditorily (Bosch & Sebastian-Galles, 1997; Mehler et al., 1988) and visually (Weikum et al., 2007), the ability to *integrate* both auditory and visual information may play a more crucial role in early language development than previously thought (Weikum et al., 2007; Yeung & Werker, 2013). Furthermore, integrating information from at least two modalities may index neural and cognitive efficiency, because integrative processes follow situation-dependent rules (Burr & Alais, 2006) and perceived simultaneity depends on compensatory strategies that realign incoming visual stimuli with incoming (relatively slower) auditory stimuli (Burr & Alais, 2006; Schroeder et al., 2008).

It is frequently reported that children and adults with neurodevelopmental disorders present with language delay, even when their subsequent language is relatively proficient, as is the case for Williams syndrome (Paterson, Brown, Gsodl, Johnson, & Karmiloff-Smith, 1999). Could this delay partly result from a deficit in integrating auditory and visual speech information? Multisensory dysfunction has already been identified in three atypically developing populations: autism spectrum disorder, dyslexia, and schizophrenia (see Wallace & Stevenson, 2014, for review), so it may present a more widespread problem in neurodevelopmental disorders. In this paper, we first discuss inter-sensory processing and speech perception—including the importance of the visual speech cue—in typically developing (TD) children. We then highlight what is currently known about these processes

in young children with neurodevelopmental disorders. Finally, we present new data on audio-visual (AV) speech perception in infants and toddlers with Down syndrome (DS), fragile X syndrome (FXS), or Williams syndrome (WS), as well as in TD controls.

Inter-sensory Processing in Typically Developing Children

Stimuli are continuously bombarding our senses. We receive far more stimulation than we can attend to or perceive. Nevertheless, we have learned to match sights and sounds, to integrate and cohere rapidly changing information from different modalities (visual, auditory, tactile, proprioceptive, and so on), in order to construct and selectively attend to unified percepts – unitary multimodal events, such as a speaking face. TD infants begin to develop this skill within the first 6 months of life, and it becomes a key building block upon which higher-level skills, such as social orienting, joint attention, and word learning, are subsequently built (Bahrick, 2010; Bahrick & Todd, 2012; Mani, Mills, & Plunkett, 2012).

The drive to cohere multisensory information into unitary representations rests on the ability to detect and process amodal information, such as temporal synchrony (Bahrick, 2010; Bahrick, Flom, & Lickliter, 2002; Bahrick, & Lickliter, 2002; Bahrick & Todd, 2012; Lewkowicz, 2000). This ability develops very early. TD infants can detect face-voice synchrony (during speech) from at least 2 months of age (Dodd, 1979; Lewkowicz, 2010; Lewkowicz, Leo, & Simion, 2010; Morrongiello, Fenwick, & Chance, 1998), although the ability to lip-read (i.e., detect spectral information common to mouth movements and speech sounds) is not mastered until some 3 months later (Kuhl & Meltzoff, 1982; Kuhl, Williams, & Meltzoff, 1991). By 6 months, TD infants can detect and use a wide range of amodal information (e.g., synchrony, rhythm, tempo, intensity), which they use to match sights and sounds, in order to organise incoming stimuli and make sense of the world around them. Additionally, according to the intersensory redundancy hypothesis (Bahrick

& Lickliter, 2000, 2002), selective attention is heightened whenever the same amodal information is concurrently and synchronously available to multiple senses (*intersensory redundancy*). This may explain why infants are drawn to social events, because these often provide the greatest amount of intersensory redundancy, making social events pop out (Bahrick, 2010). It also means that intersensory processing may play a key role during early ontogenesis and have cascading effects on the emergence of higher-level socio-cognitive processes as well as social behaviours such as social orienting and speech perception.

Speech Perception in Typically Developing Children

Speech perception is often multimodal (Campbell, 1996, 2008; Kushnerenko, Teinonen, Volein, & Csibra, 2008). Lip movements usually accompany speech sounds; watching lip movements influences auditory (speech) perception (Alsius, Navarra, Campbell, & Soto-Faraco, 2005; Burnham & Dodd, 2004; Kushnerenko et al., 2008; McGurk & MacDonald, 1976; Rosenblum, Schmuckler, & Johnson, 1997). For example, when infants are shown a visual display of two side-by-side talking faces and hear a synchronously-presented sound, they spontaneously look longer at the face that matches the sound (the *congruent* face) than the face that does not (the *incongruent* face) (Kuhl & Meltzoff, 1982, 1984; Kuhl, Williams, & Meltzoff, 1991; MacKain, Studdert-Kennedy, Spieker, & Stern, 1983; Patterson & Werker, 1999, 2003). Visual information may be especially helpful for speech perception under noisy conditions (Sumby & Pollack, 1954), thereby contributing to infants' learning of auditory phoneme categories (Teinonen, Aslin, Alku, & Csibra, 2008).[1]

Auditory-Visual Integration in Down, Fragile X, and Williams syndromes

[1] Although AV integration *facilitates* speech perception, it is not a necessary requirement. For instance, despite there being basic differences between blind and sighted children (e.g., Andersen, Dunlea, & Kekelis, 1984), blind children are of course capable of acquiring language (Perez-Pereira & Conti-Ramsden, 2013).

Although AV speech integration facilitates word learning (Teinonen, Aslin, Alku, & Csibra, 2008), it has rarely been investigated in neurodevelopmental disorders such as Down syndrome (DS), Fragile X syndrome (FXS), or Williams syndrome (WS), and never previously in infants. One study focused on older children and adults with WS and demonstrated impairments in a speech-reading condition, even though the participants performed as well as Chronological Age (CA)-matched controls when they were instructed to repeat vowel-consonant-vowel (VCV) syllables presented auditorily (Böhning, Campbell, & Karmiloff-Smith, 2002). Although the individuals with WS were compared with CA-, but not Mental Age (MA)-, matched controls, the study suggests that visual information processing rather than AV integration per se is impaired in WS. Yet the data did not lend themselves to reject the hypothesis that the actual integration of audio-visual stimuli is impaired in WS (Böhning, Campbell, & Karmiloff-Smith, 2002). Moreover, Massaro (1987, 1998) argues that AV speech integration is likely to be atypical in any population if either auditory or visual processing is impaired.

Even less is known about AV integration in DS and FXS. However, individuals with DS or FXS often show the opposite pattern of auditory and visual processing to those with WS, i.e., auditory information processing is frequently more impaired than visual information processing in FXS (Van der Molen et al., 2012; see also Scerif, Longhi, Cole, Karmiloff-Smith, & Cornish, 2012, for evidence of poorer auditory attention [relative to visual attention] and the suggestion that multimodal information processing is atypical in FXS). Additionally, auditory memory is worse than visual memory in DS (e.g., Marcell & Armstrong, 1982), which indicates that, in at least some tasks, individuals rely more on one modality than another. Again, it is unlikely that AV speech integration is typical in these populations if auditory processing is impaired relative to visual processing. Nevertheless, *pace* Massaro (1987, 1998), we do not actually know whether the ability to integrate

auditory and visual speech information is compromised in the three neurodevelopmental disorders focused on in this study and particularly whether such a deficit is already manifest in very early development.

The Current Task

In sum, children and adults with DS, FXS, and WS all present with language delay (Rice, Warren, & Betz, 2005), but it remains unclear whether deficits in AV integration contribute to the delay in these populations. Impairment would affect their early abilities to perceive speech (Kushnerenko et al., 2008), learn phoneme boundaries (Teinonen et al., 2008), and ultimately acquire language. The current study seeks to ascertain whether AV speech integration is atypical in infants and toddlers with DS, FXS, and/or WS. We address this question by measuring AV speech integration in infants and toddlers with the three genetic syndromes, and comparing the data with CA- and MA-matched TD controls.

AV speech integration can be assessed by testing for the McGurk effect (McGurk & MacDonald, 1976). The McGurk effect is a perceptual phenomenon that occurs when the auditory component of one sound (e.g., hearing /ba/) is paired with the visual component of another sound (e.g., seeing /ga/), leading to the perception of a third sound (e.g., perceiving /da/ or /tha/) (McGurk & MacDonald, 1976; see also Basu Mallick, Magnotti, & Beauchamp, 2015). This effect has been demonstrated as early as 4.5 months of age in TD infants (Burnham & Dodd, 2004; Rosenblum, Schmuckler, & Johnson, 1997; but see Desjardins & Werker, 2004, for evidence that AV speech integration is neither strong nor consistent in infants; see also Baart, Vroomen, Shaw, and Bortfeld, 2014, for evidence that infants may rely on different cross-modal cues than adults when matching auditory and visual speech).

In case it turns out to be true that AV speech integration (the McGurk effect) is impaired in children with neurodevelopmental disorders, we decided also to measure a

more basic type of AV speech "integration", namely, intermodal matching. This can be assessed by testing for sensitivity to matching vs. mismatching (co-occurring) audio-visual speech stimuli (e.g., Guiraud et al., 2012). Specifically, a sound (e.g., /ga/) is presented to the participant together with two faces; one face silently mouths a sound that matches the auditory stimulus (/ga/), while the other face silently mouths a non-matching sound (e.g., /ba/). (This may also lead to the perception of a third sound – e.g., perceiving /bga/. This is a combination response rather than a fusion response [Omata & Mogi, 2008].) Although there is no evidence that infants are capable of integrating auditory and visual speech information into a single percept before 4.5 months of age (see above), Patterson and Werker (2003) showed that infants as young as 2 months are sensitive to (co-occurring) matching/mismatching audio-visual speech stimuli.

In other words, because TD infants as young as 2 months are sensitive to (co-occurring) matching/mismatching audio-visual speech stimuli, we believe it is highly likely that 16-month-olds will also be sensitive to such stimuli even if they have DS, FXS, or WS. Integrating information from two different modalities to form an illusory percept is, however, a more sophisticated kind of AV integration that happens later in infancy (see Nardini, Bedford, & Mareschal, 2010, for evidence that children do not experience fusion in all cases until at least 8-12 years of age).

Therefore, it was hypothesised that our TD controls would demonstrate good AV integration skills by:

(1) looking longer at the congruent face than the incongruent face in a 'Mismatch' condition, in which an auditory /ga/ is matched with a visual /ga/ for the congruent face, and a visual /ba/ (to produce the non-English syllabic /bga/) for the incongruent face (thus demonstrating intermodal matching [see Figure 1]); while

(2) failing to discriminate between the congruent face and the incongruent face in a 'Fusion' condition (i.e., by looking at both faces equally), which is when an auditory /ba/ is matched with a visual /ga/ to produce the normal English syllabic percept /da/ for the incongruent face (and the English syllabic /ba/ for the congruent face) (Figure 1). In other words, in the Fusion condition, the TD controls should perceive conflicting percepts in both the congruent face and incongruent face (see Figure 2 for a hypothetical depiction of the predictions).

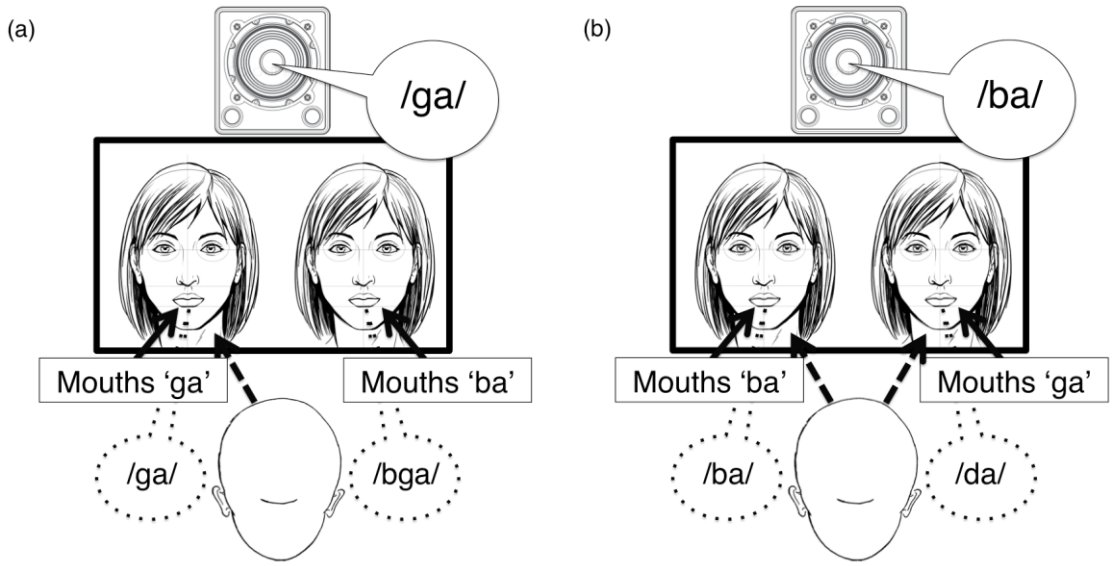


Figure 1. (a) In the Mismatch condition, participants hear /ga/ while one face mouths 'ga' and the other mouths 'ba'. When participants see the face mouthing a 'ga', they perceive a /ga/ because what they are seeing is congruent with what they are hearing. However, when participants see the 'ba', they perceive a mismatch (e.g., a combinatorial percept such as /bga/). This is because what they are seeing is incongruent with what they are hearing. Infants with good intermodal matching ability can discriminate between these kinds of congruent and incongruent faces; they look longer at the congruent face. (b) However, the sound /ba/ and sight 'ga' (the Fusion condition) fuse into the syllabic percept /da/ (the *McGurk effect*). Participants with a strong AV speech integration ability should demonstrate a McGurk effect and thus fail to discriminate between the faces because there is nothing to distinguish one syllabic percept (e.g., /da/) from the other (/ba/).

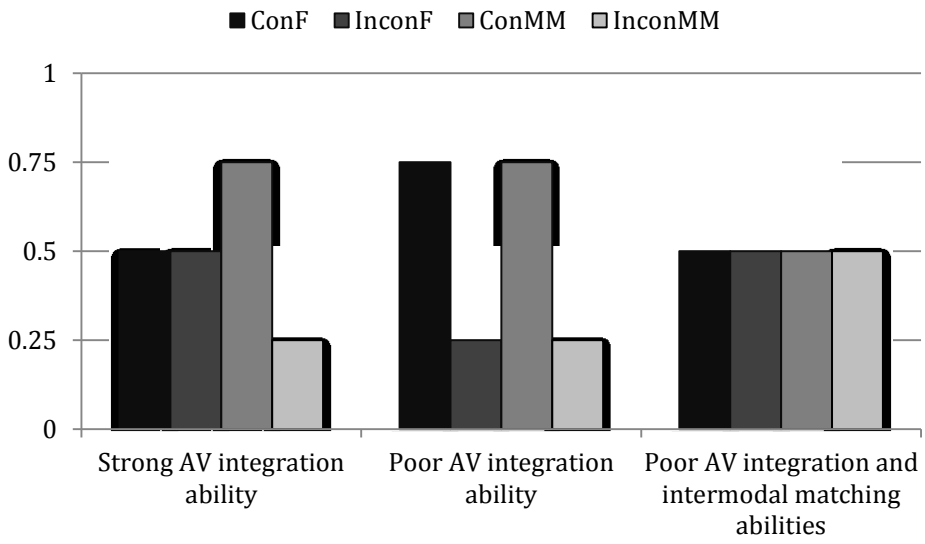


Figure 2. A hypothetical depiction of the predictions: Participants with a strong AV speech integration ability should discriminate between the congruent face and the incongruent face in the Mismatch condition but not in the Fusion condition; participants with poor AV speech integration ability should fail to show a McGurk effect and thus discriminate between congruent and incongruent faces *irrespective of Condition* (fusion, mismatch); participants with poor intermodal matching ability should fail to discriminate between faces.

The atypically developing groups were expected to discriminate between the faces in both conditions. This is because we hypothesized that they have impaired AV integration and not necessarily the inability to detect incongruence between AV information (an ability that develops early in typical development). No a priori predictions were made with respect to potential differences between the DS, FXS, and WS groups, although we expected this cross-syndrome design to yield interesting subtle differences. See Table 1 for a summary of the predictions.

Table 1.
Summary of predictions by Condition and Group: Will the children discriminate between the congruent and incongruent faces (Yes or No)?

Group	Condition	
	Fusion	Mismatch

TD controls	N	Y
DS	Y	Y
FXS	Y	Y
WS	Y	Y

Note: Y = Yes, the children will discriminate between the faces.
N = No, the children will not discriminate between the faces.

Methods

Participants

A total of 95 atypically developing infants and toddlers with one of the three neurodevelopmental disorders was tested: 22 infants and 21 toddlers with DS (*N*=43), 14 toddlers with FXS (too few infants with FXS were available for testing), and 12 infants and 26 toddlers with WS (*N*=38). The participants had been clinically diagnosed and/or genetically tested respectively for full trisomy 21, mutation of the FMR1 gene, or deletion of the ELN gene. Data collected from these children were compared with data from 25 typically developing (TD) controls. Data from these controls were made available through the British Autism Study of Infant Siblings (BASIS, www.basisnetwork.org; NHS NRES London REC 08/H0718/76).

Because children with DS, FXS, or WS have a mental age (MA) of approximately half their chronological age (CA), data from the TD control group were compared with data from MA-matched groups as well as CA-matched groups. For the purpose of comparison, participants' mental ages were obtained using the Mullen Scales of Infant Learning (MSEL; Mullen, 1995). Data from one participant with WS were excluded from all analyses, because the 37-month-old had an unusually high MA (31.25 months) that was 2.76 standard deviations (SD) above the WS group mean.

Table 2 displays the mean CA for each CA-matched group (*TD controls*, *DS infants*, *WS infants*). A one-way ANOVA shows that the CA-matched groups did not

significantly differ on CA, $F_{2,56} = 1.60$, $p = .212$. Because the distribution of CA data in the WS and TD control groups looked slightly bimodal, an Independent-samples Kruskal-Wallis test was also carried out to confirm the results of the ANOVA. This non-parametric test also yielded no significant difference between the groups on CA, $H(2) = 1.32$, $p = .516$.

Table 2.
Mean chronological age (CA), mental age (MA), and verbal mental age (VMA) for each group.

Group	*N*	CA in months *(SD)*	MA in months *(SD)*	VMA in months *(SD)*
TD controls	25	15.48 *(0.82)*	16.60 *(2.50)*	16.10 *(3.73)*
DS infants[a]	22	16.23 *(1.81)*	8.46 *(2.45)*	7.19 *(2.59)*
WS infants[a]	12	16.13 *(2.00)*	8.71 *(1.89)*	7.13 *(2.26)*
DS toddlers[b]	21	28.83 *(6.86)*	15.86 *(4.52)*	15.02 *(5.57)*
FXS toddlers[b]	14	34.39 *(8.32)*	15.34 *(4.42)*	13.21 *(5.95)*
WS toddlers[b,c]	25	30.14 *(8.23)*	16.11 *(4.53)*	15.63 *(5.76)*

[a] These two groups were CA-matched to the TD control group

[b] These three groups were MA- and VMA-matched to the TD control group

[c] This table does not include the participant with WS who was excluded from the analyses for having an unusually high MA (see main text).

MA data were normally distributed. A one-way ANOVA revealed that MA did not significantly differ across the MA-matched groups (*TD controls*, *DS toddlers*, *FXS toddlers, WS toddlers*), $F_{3,80} = 0.32$, $p = .809$ (Table 2). Because infant scanning of AV speech stimuli is related to language development (D'Souza et al., 2015), we checked whether the MA-matched groups were also matched on verbal MA. The verbal MA data were not normally distributed, so a non-parametric test was used. Verbal MA did not significantly differ across groups, $H(3) = 2.52$, $p = .472$ (Table 2).

Design

An intermodal preferential looking paradigm (IPLP: Golinkoff, Hirsh-Pasek, Cauley, & Gordon, 1987) was used to investigate the McGurk effect (McGurk & MacDonald, 1976)—and thus AV integration—and intermodal matching in the infants and toddlers. The design was adapted from Kushnerenko et al. (2008) and Tomalski et al.

(2013). Adult female heads, with moving lips, were presented on a screen, with loudspeakers placed behind it. The position of the heads was pseudorandomised across trials and participants. Four trials of 15 repetitions of moving lips and sounds were presented. The four trials were divided into two Fusion and two Mismatch trials. As mentioned above, in adults the visual /ga/ and auditory /ba/ (the Fusion condition) integrate to produce an English syllabic percept /da/, while the visual /ba/ and auditory /ga/ (the Mismatch condition) integrate to produce a non-English syllabic percept /bga/ (McGurk & MacDonald, 1976).

The order of presentation of the trials was counterbalanced across infants, so that the same number of infants saw the Fusion condition first and second, and the Congruent head on the left and right sides of the screen; the same applied to the Mismatch condition. In each trial, the participants were presented with auditory-visual (AV) stimuli: two talking heads (one on each side of the screen) and a sound (see Figure 3). The two Fusion trials were as follows: in one trial, the left head mouthed the syllable /ba/ and the right head mouthed the syllable /ga/, while the sound /ba/ was simultaneously heard (Figure 1b); in the other (*Fusion*) trial, the left head mouthed the syllable /ga/ and the right head mouthed the syllable /ba/, while the sound /ba/ was simultaneously heard. The two Mismatch trials were as follows: in one trial, the left head mouthed the syllable /ba/ and the right head mouthed the syllable /ga/, while the sound /ga/ was simultaneously heard; in the other (*Mismatch*) trial, the left head mouthed the syllable /ga/ and the right head mouthed the syllable /ba/, while the sound /ga/ was simultaneously heard (Figure 1a). Thus, in each trial, one speaking head is *congruent* (i.e., the visual stimuli match the auditory stimuli) and the other speaking head is *incongruent*, and one pair should give rise to a fused percept and the other to a mismatch.

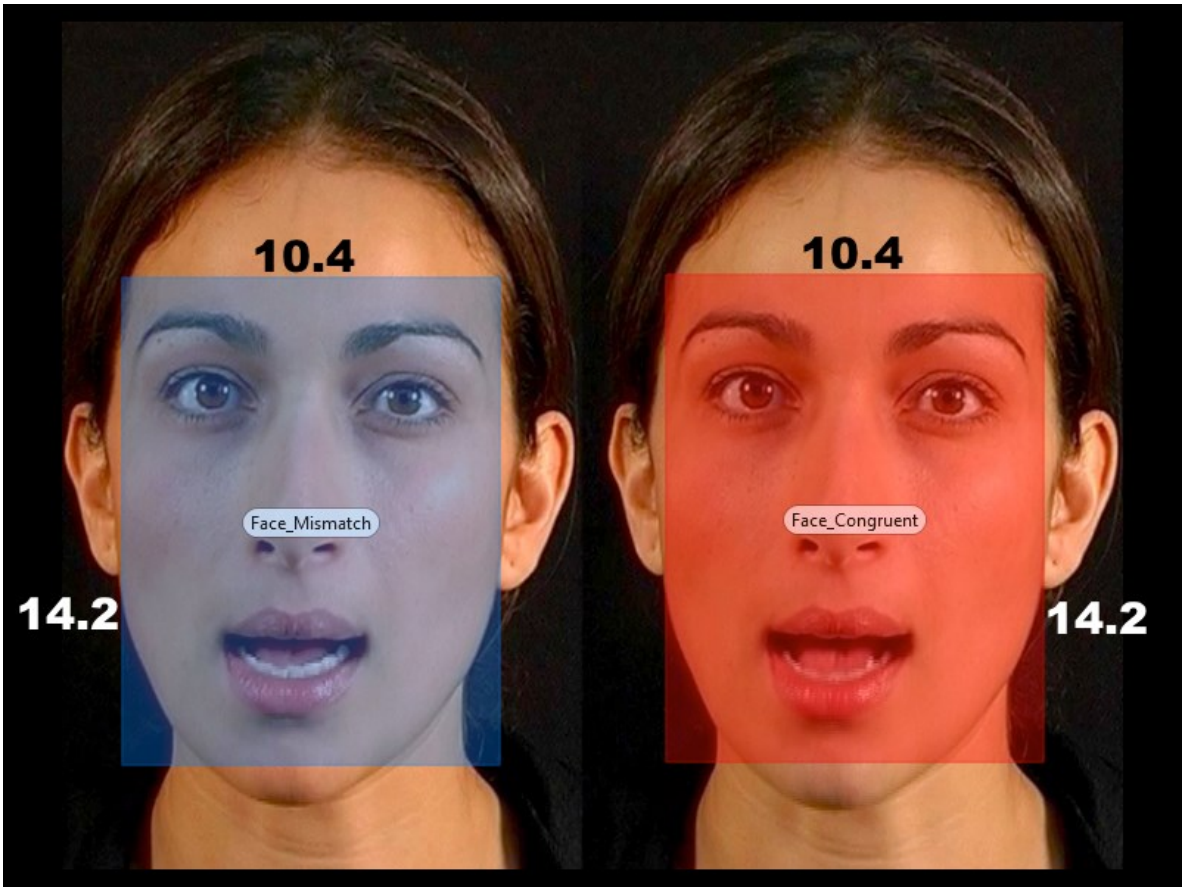


Figure 3. An example of the visual stimuli used in the experiment, including the positioning and size in visual angle of the Area-Of-Interest.

Materials

The same stimuli were used as in Kushnerenko et al. (2008; see also Guiraud et al. [2012] and Tomalski et al. [2013]). The stimuli consisted of four 12-second video clips of two speaking faces, presented side-by-side, against a black background (see Figure 1 for an example of the visual stimuli, including the positioning and size in visual angle of the Area-of-Interest). The faces were of the same female native English speaker. One face mouthed the /ba/ syllable, while the other face mouthed the /ga/ syllable. The faces on the monitor were approximately life size. In two of the four video clips, the sound /ba/ was played, while in the other two video clips the sound /ga/ was played. The video clips and stereo

soundtracks were digitised at a rate of 25 frames per second and 44.1 kHz with 16-bit resolution.

Each 12s video clip started with lips fully closed, with the face being silent for the first nine frames (360 ms). The subsequent voiced section lasted for seven or eight frames, followed by two to three frames of the mouth closing. Thus, sound onset began 360 ms after stimulus onset, and the voiced section lasted for 280-320 ms. The mouth opening for the visual /ga/ stimulus started about 260ms prior to sound onset. For the visual /ba/, it started simultaneously with sound onset, with the lips pressing together around 280 ms prior to sound onset. Total duration of the stimulus (i.e., one mouth movement and a simultaneous sound) was 760 ms and stimulus onset asynchrony (SOA) was 760 ms, with each video clip of 15 repetitions being 12s long.

In other words, there were four video clips (two conditions). Each video clip lasted 12 seconds. Each video clip included 15 repetitions of mouth movements and a simultaneous sound (/ba/ or /ga/). Each repetition (which included the opening and closing of the mouth) lasted for 760 ms. Each repetition included the sound (/ba/ or /ga/), which lasted for 280 ms. Incongruent face stimuli were created by dubbing the auditory /ba/ onto the visual /ga/ and vice versa. The consonantal burst in the audio file was aligned with the consonantal burst in the video file.

Kushnerenko and colleagues (2008) pre-tested the stimuli to five native adult English speakers. Four of them reported hearing /da/ or /ta/ in the Fusion condition and either /bga/ or mismatched audiovisual input in the Mismatch condition, and one of adults reported hearing only the auditory component in both conditions.

A Tobii T120 remote eye tracker (Tobii Technology AB) was used to capture infants'/toddlers' moment-to-moment point of gaze at a sampling rate of 120 Hz, with measurement accuracy of about 0.5°. The visual stimuli were presented on a 34 x 27 cm

TFT liquid crystal display monitor, with a resolution of 1280 x 1024 pixels and a response rate of 4 ms. The tracking equipment and stimulus presentation were controlled using Tobii Studio 2.1.14. A camera mounted directly above the horizontal midpoint of the screen was used to monitor and record infant behaviour. The auditory stimuli were delivered via two speakers positioned behind the display monitor and facing the participant.

Procedure

The same procedure was used as in Guiraud et al. (2012) and Kushnerenko et al. (2008) with TD children. Infants sat on their parent's lap, in a dimly lit featureless room, facing the stimulus-presentation screen with their eyes at a distance of approximately 60 cm from the screen. The experimenter sat behind a curtain and observed the infant, using Tobii Studio LiveViewer via a camera that was positioned centrally above the screen. The infants' eye movements were recorded using Tobii Studio 2.1.14. Caregivers were asked to close their eyes during the experiment. Calibration was carried out using 5 points: one in each corner of the screen and one in the centre of the screen. Before each trial, a colour animation and interesting sounds were played to attract the infant's attention to the centre of the screen. Once the child's attention was focused on the screen centre, the attention grabber was terminated and the trial was started simultaneously. The stimuli were presented in a pseudorandom order and counterbalanced across participants. A distractor separated each trial. The whole procedure lasted less than 10 minutes.

Analysis

For each participant, the quality of recording was measured as a percentage (the number of eye tracking samples that were correctly identified, divided by the number of attempts, so 50% means that one eye was found for the full recording or that both eyes were found for half the time. The eyes cannot be detected when a participant is looking away from the screen; this will result in a lower percentage). The quality of recording was at least

25% in all participants. The quality of recordings did not significantly differ across CA- or MA-matched groups, $H(2) = 4.76$, $p = .093$, $F_{3,81} = 0.90$, $p = .448$, respectively (see Table 3). A non-parametric test was used for the CA-comparison because the data in the WS infant group had a continuous uniform (rather than a normal) distribution.

Table 3.

Quality of data for each group.

Group	*N*	Mean (%)	*SD*
TD controls	25	82.2	12.7
DS infants[a]	22	80.1	14.3
WS infants[a]	12	63.4	25.1
DS toddlers[b]	21	75.3	13.1
FXS toddlers[b]	14	80.1	12.8
WS toddlers[b,c]	25	77.4	19.4

[a] These two groups were CA-matched to the TD control group

[b] These three groups were MA-matched to the TD control group

[c] This table does not include the participant with WS who was excluded from the analyses for having an unusually high MA (see main text).

An Area-Of-Interest (AOI) was delineated around the face (Figure 2). This was defined before data were collected (Guiraud et al., 2012; Tomalski et al., 2013). Fixation measures were calculated off-line using Tobii Studio and Tobii fixation filter (Tobii Inc.).

Because different measures tap into different cognitive and neural processes (Cohen, 1972; Schiller & Tehovnik, 2001), two measures were used: proportion of looking time (fixation duration) and first fixation.

(1) *Proportion of time looking* (PTL) is the total amount of time spent looking at the target stimulus (e.g., the congruent face) as a proportion of the total amount of time spent looking at both the target and the distractor stimulus (i.e., the congruent face / [the incongruent face + the

congruent face]). *Total amount of time spent looking* is the total duration of all fixations within an AOI. PTL is a classic measure used in studies of infant perception, cognition, and development (Aslin, 2007).

(2) *First fixation* is defined as the first fixation on an AOI in a trial. We included this measure because, if a participant could easily discriminate the different mouth movements between the two identical faces, then s/he might show an effect within the very first few fixations (e.g., look longer at one face at first), but not over the duration of the whole trial (e.g., s/he might subsequently look at the other face simply out of boredom or curiosity and, thus, at both faces more or less evenly, despite a first fixation at the critical face). (See Rupp & Wallen, 2007, for an example of this measure being used in an eye-tracking study.)

All data were tested for normality. Data values that were above or below 2 standard deviations from the group mean were judged *a priori* to be outliers, and hence removed from the analysis (unless the data were transformed – see below). This decision was based on the literature (Field, 2013). If the data were non-normal (i.e., if $Z_{Skewness} > 2$ and Kolmogorov-Smirnov $p < .05$), then they were transformed (log 10 to correct for positive skew, reflected log 10 for negative skew, or arcsine for proportions, as appropriate). If the transformed data were non-normal (e.g., bimodal), then the untransformed data were analysed using the appropriate non-parametric test (e.g., Mann-Whitney).

Results

Analyses of *fixation duration* yielded only non-significant results. For instance, there was no Face type (congruent, incongruent) x Condition (fusion, mismatch) x Group

interaction in either the CA-matched comparison or the MA-matched comparison ($F_{2, 45}$ = 1.14, p = .328, $F_{3, 70}$ = 1.20, p = .318, respectively).

However, analyses of *first fixation* were—as predicted—more revealing:

Control data

Ninety-one of 100 trials were valid. We ran a 2 (Face: congruent, incongruent) x 2 (Condition: fusion, mismatch) repeated measures ANOVA. TD controls were more likely to fixate first on a congruent face than an incongruent face (53 vs. 36), $F_{1, 22}$ = 5.95, p = .023, η = .21. There was no main effect of Condition, $F_{1, 22}$ = 0.32, p = .575. However, there was an interaction effect, $F_{1, 22}$ = 4.43, p = .047, η = .17 (see Figure 4).

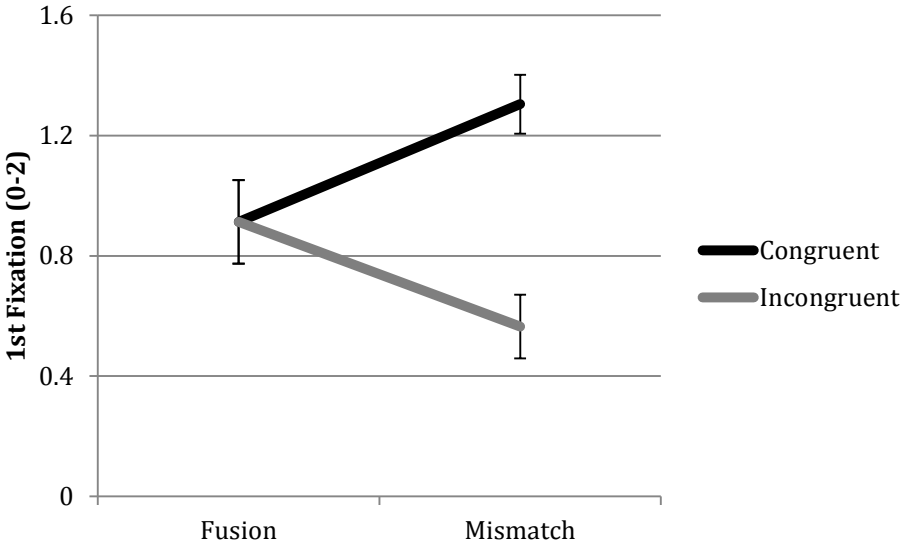


Figure 4. In TD controls, first fixation differentiated between the congruent and incongruent faces in the Mismatch, but not Fusion, condition, as predicted. Error bars represent ±1 standard error of the mean.

Two (post hoc) paired-samples *t*-tests were carried out to break down the interaction effect. Bonferroni corrections were applied. Therefore, all effects are reported at a .025 level of significance. In the Fusion condition, as predicted TD controls' first fixation did

not differentiate between congruent and incongruent faces (23 vs. 22), $t(24) = 0.16$, $p = .914$. However, in the Mismatch condition, TD controls were significantly more likely to direct a first fixation towards the congruent face than the incongruent face (30 vs. 14), $t(22) = 3.87$, $p = .001$, $r = .64$.

Although $Z_{Skewness}$ was less than ±2 for each face and condition, the data looked non-normal and thus non-parametric Wilcoxon Signed Rank Tests were carried out to confirm the above findings. Whereas first fixation did not differentiate between congruent and incongruent faces in the Fusion condition (positive differences = 6, negative differences = 7, number of ties = 12), $T = 44$, $Z = 0.11$, $p = .914$, it did in the Mismatch condition (positive differences = 0, negative differences = 10, number of ties = 13), $T = 0.00$, $Z = 2.92$, $p = .004$. See Figure 5 for overall frequency data and Figure 6 for information on individual scores.

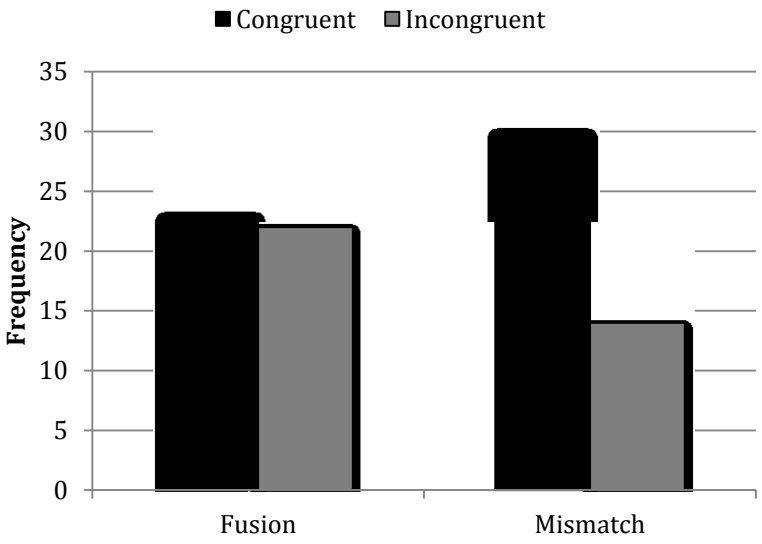


Figure 5. In the Fusion condition, number of first fixations was similar across faces (congruent, incongruent) in TD controls. In the Mismatch condition, the TD infants were more likely to direct their attention towards the congruent face than to the incongruence face.

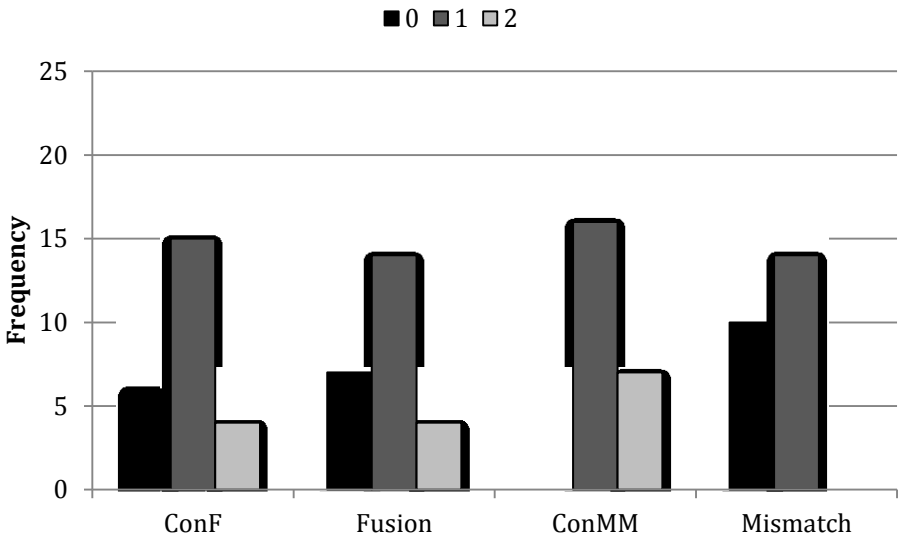
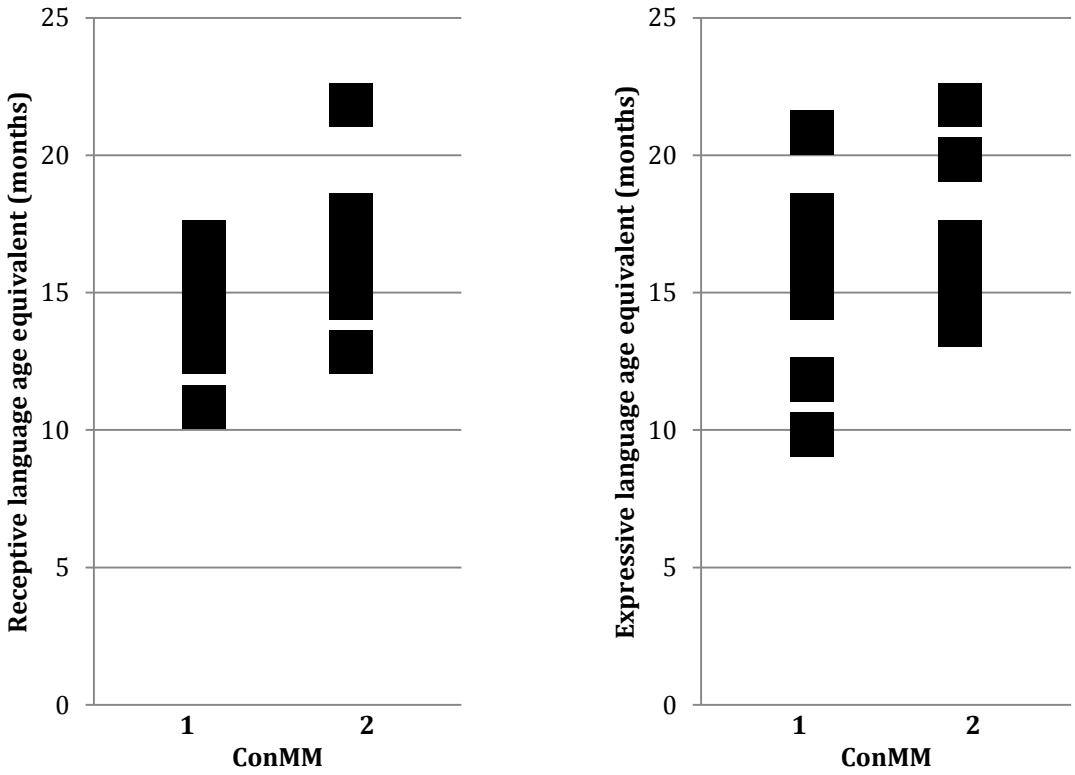


Figure 6. Frequency of individual scores: the legend represents 'score' (i.e., over two trials, 0 = no first look, 1 = one first look, 2 = two first looks), In the Fusion condition, number of first fixations was similar across faces (congruent, incongruent) in TD controls. In the Mismatch condition, all TD infants made at least one first look to the congruent face, and no TD infant made two first looks to the incongruence face.

Post hoc analyses: The relationship between AV speech integration and language ability. Because first fixation differentiated between faces in the Mismatch condition, and infant scanning of AV speech stimuli is related to language ability (D'Souza et al., 2015), we analysed whether first fixation was predictive of language ability in the TD controls. We carried out a regression analysis with first fixations to the congruent face in the Mismatch condition as the predictor, and with receptive or expressive language ability (measured using the Mullen Scales of Early Learning) as the outcome variable.

Receptive language. The regression model was significant ($F_{1,19} = 4.46$, $p = .048$; $B = 2.43$, $SE = 1.15$, Beta = 0.44, CI = 0.02, 4.83; see Figure 7a). First fixations to the congruent face accounted for 19% of the variance in receptive language ability (adjusted $R^2 = .15$).

Expressive language. Although visual inspection of the data reveal a similar pattern to the one for receptive language (compare Figure 7b with Figure 7a), the regression model was not significant when expressive language was the outcome variable ($F_{1,20} = 1.47$, $p = .240$).



(a) (b)

Figure 7. In the Mismatch condition, all TD infants made either one or two first looks to the congruent face. (a) Infants who made two looks had a significantly higher receptive language score than those who made only one look. For example, whereas one infant who made two looks had a receptive language score equivalent to a 22-month-old, one infant who made only one look had a score equivalent to an 11-month-old. (b) Infants who made two looks to the congruent face also had a higher expressive language score than those who made only one look, but this difference was not statistically significant.

CA-comparison (N.B. no infants with FXS participated)

One infant with WS failed to provide valid data. The remaining children provided 104 valid trials (in DS, 74 of 88 trials were valid [84%]; in WS, 30 of 44 were valid

[68%]). Some of the DS and WS data were non-normal ($Z_{skewness} > \pm 2$, even when logarithmically transformed). Therefore, non-parametric Wilcoxon Signed Rank Tests were carried out (on the untransformed data).

In contrast to TD controls, infants with WS were significantly more likely to land a first fixation on the congruent face than on the incongruent face in the Fusion condition (positive differences = 0, negative differences = 5, number of ties = 6), $T = 0$, $Z = 2.24$, $p = .025$, $r = .67$ but, unlike the TD controls, not in the Mismatch condition (positive differences = 2, negative differences = 5, number of ties = 4), $T = 6$, $Z = 1.41$, $p = .160$. Due to the paucity of data in the WS infant group, however, this finding should be interpreted with caution. In the DS infant group, first fixation failed to differentiate between faces in either condition, $T = 35$, $Z = 0.80$, $p = .426$ (Fusion: positive differences = 5, negative differences = 5, number of ties = 10), $T = 43$, $Z = 1.62$, $p = .105$ (Mismatch: positive differences = 6, negative differences = 4, number of ties = 11). See Figure 8. See Figure 9 for a more nuanced breakdown of the WS data.

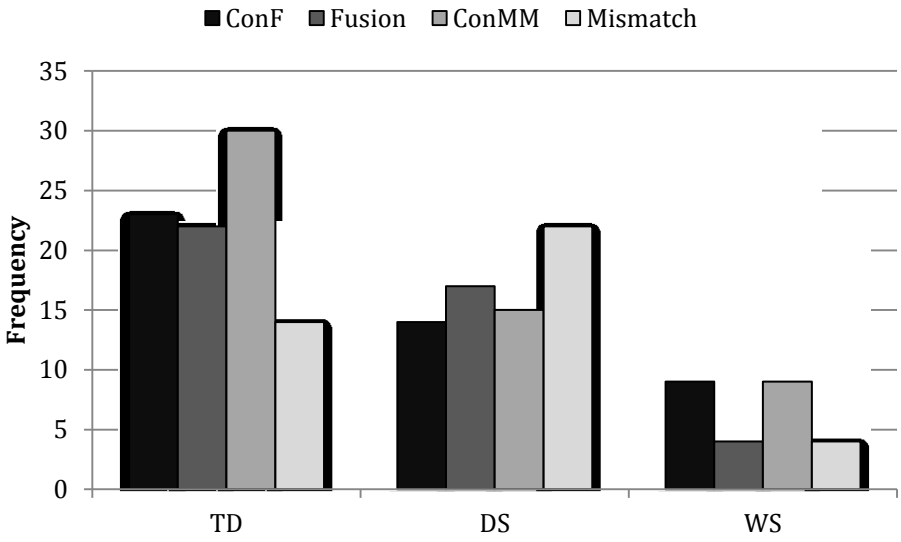


Figure 8. In TD controls, first fixation differentiated between the congruent and incongruent faces in the Mismatch, but not Fusion, condition. In WS, the opposite was true (but see Figure 9). In DS, first fixation did not differentiate between faces in either of the conditions. ConF = congruent face in the Fusion condition. InconF = incongruent face in

the Fusion condition. ConMM = congruent face in the Mismatch condition. InconMM = incongruent face in the Mismatch condition. Error bars represent ±1 standard error of the mean.

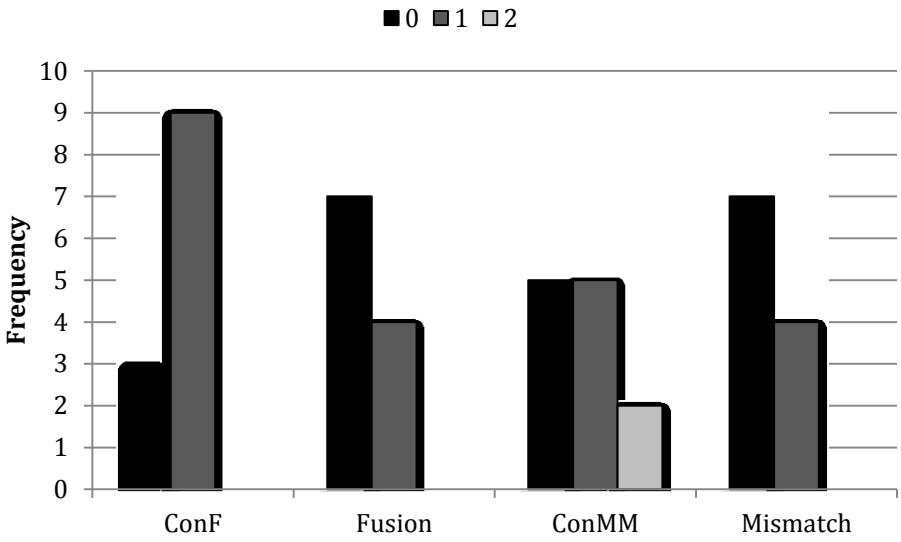


Figure 9. Frequency of individual scores in the Williams syndrome (WS) group; the legend represents 'score' (i.e., over two trials, 0 = no first look, 1 = one first look, 2 = two first looks). Infants with WS were significantly more likely to land a first fixation on the congruent face than on the incongruent face in the Fusion condition (9 vs. 4). A similar (but not statistically significant) pattern was found in the Mismatch condition (9 vs. 4). The reason that the second result was non-significant is because some of the infants provided data for only one of the two trials, and whereas five of these infants looked only at the congruent face, two of them looked only at the incongruent face. Due to the paucity of data in the WS infant group, this finding should be interpreted with caution.

MA-comparison (N.B. including participants with FXS)

All participants provided at least some valid data. In DS, 79 of 84 trials were valid (94%). In FXS, 53 of 56 trials were valid (95%). In WS, 83 of 100 trials were valid. Some of the FXS and WS data were non-normal ($Z_{skewness} > \pm 2$, even when logarithmically transformed). Therefore, non-parametric Wilcoxon Signed Rank Tests were carried out (on the untransformed data).

For all groups (except the TD control group; see above), first fixation failed to differentiate between the faces in either of the two conditions (Fusion: $T = 21.50$, $Z = -0.12$, $p = .902$ (DS), $T = 0.00$, $Z = 0.00$, $p = 1.000$ [all ties] (FXS), $T = 50.00$, $Z = -0.16$, $p = .871$ (WS); Mismatch: $T = 19.50$, $Z = 0.22$, $p = .826$ (DS), $T = 16.50$, $Z = 0.44$, $p = .660$ (FXS), $T = 3.50$, $Z = -1.63$, $p = .102$ (WS)). See Figure 10.

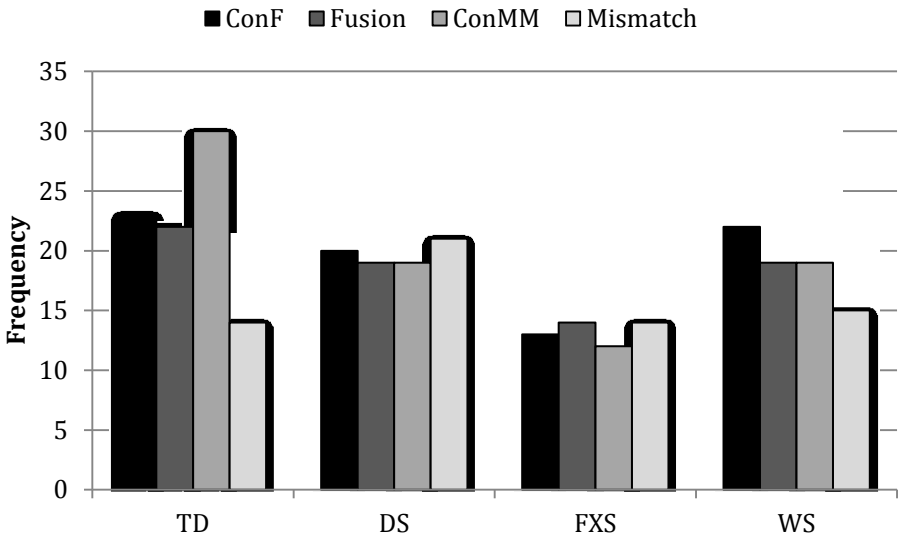


Figure 10. In TD MA-matched controls, first fixation differentiated between the congruent and incongruent faces in the Mismatch, but not Fusion, condition. In DS, FXS, and WS, first fixation did not differentiate between faces in either of the conditions. ConF = congruent face in the Fusion condition. InconF = incongruent face in the Fusion condition. ConMM = congruent face in the Mismatch condition. InconMM = incongruent face in the Mismatch condition. Error bars represent ±1 standard error of the mean.

Other Areas-Of-Interest (AOIs)

We also delineated AOIs around the eyes and mouth and analysed proportion of time looking at the eyes and mouth AOIs, as reported in Guiraud et al. (2012). However, the fixation data were uninformative, whether comparing face AOIs or eyes and/or mouth AOIs.

Post hoc analyses: The relationship between AV speech integration and language ability in DS, FXS, and WS.

The regression analyses carried out on data from the TD control group (see above) were also carried out on data from the atypically developing groups. The AV speech integration data failed to predict language ability in any of the atypically developing groups, though there was a trend in toddlers with DS for expressive language (DS infants: $F_{1,14} = 0.11$, $p = .746$ (receptive), $F_{1,14} = 1.67$, $p = .217$ (expressive); WS infants: $F_{1,10} = 0.18$, $p = .684$ (receptive), $F_{1,10} < 0.01$, $p = .949$ (expressive); DS toddlers: $F_{1,18} = 0.96$, $p = .340$ (receptive), $F_{1,18} = 4.05$, $p = .059$ (expressive); FXS toddlers: $F_{1,12} = 1.19$, $p = .296$ (receptive), $F_{1,12} < 0.01$, $p = .946$ (expressive); WS toddlers: $F_{1,20} = 0.86$, $p = .364$ (receptive), $F_{1,20} < 0.01$, $p = .929$ (expressive).

Discussion

The aim of the present study was to investigate AV speech perception in three neurodevelopmental disorders: Down syndrome, fragile X syndrome, and Williams syndrome. Because children with DS or FXS have problems with auditory processing whereas those with WS have difficulty with visual processing, we hypothesised that all three groups would be impaired in AV integration. It was therefore predicted that they would fail to show a McGurk effect. In contrast, the TD controls were expected to demonstrate AV integration by making a discrimination between the faces in the Mismatch condition, but not in the Fusion condition. The TD controls showed the expected outcome; this was not the case for the neurodevelopmental disorders. The TD controls did indeed discriminate between the faces with their *first fixation* in the Mismatch condition but, as predicted, not in the Fusion condition, which suggests that they (1) perceived incongruent, non-fusible AV speech information and (2) may have integrated fusible AV speech

information. Furthermore, the ability to perceive incongruent AV speech information was linked to language ability in TD. By contrast, the CA- and MA-matched atypically developing groups failed to make a distinction in the Mismatch condition, which hints at poor detection of incongruence between AV information (though data were sparse in the CA-matched WS group and there was a trend [$p = .102$] in the MA-matched WS group). Moreover, the (CA-matched) infants with WS made a distinction between congruent and incongruent faces in the Fusion condition. This suggests that, unlike TD controls, infants with WS cannot integrate fusible AV speech information.

Do these data indicate poor AV integration in *all* the atypically developing groups? Although the descriptive data hint that children with Williams syndrome have poor AV integration ability while children with DS and FXS have poor intermodal matching ability (compare Figures 8 and 10 with Figure 2), we cannot draw this strong conclusion because there are no technical grounds for accepting null results. Failure to reject the null hypothesis suggests that either the null hypothesis is true or it is false but we do not have sufficient power to reject it. However, because the TD infants discriminated between the two faces in the Mismatch condition (note the large effect size: $r = .64$), we can firmly conclude that they perceived incongruent, non-fusible AV speech information. If they could demonstrate this, then why would they fail to discriminate between the two faces in the Fusion condition? The best explanation is that they perceived conflicting English syllabic percepts as a result of automatically integrating the AV speech information. This is consistent with findings in the previous literature. For instance, Guiraud et al. (2012) reported a similar pattern of results in a different group of TD infants using the same paradigm as the one used in the current study. Moreover, the (CA-matched) infants with WS discriminated between the congruent and incongruent faces in the Fusion condition (again, note the large effect size: $r = .67$). Thus, we suggest that, unlike the TD controls, the

infants with WS failed to integrate AV speech information. At the very least, we conclude that looking patterns in WS are unlike those observed in TD controls and are thus atypical. However, although our data suggest that intermodal matching is also atypical in DS and FXS, further evidence is required before we can confidently draw such a conclusion.

Why would AV speech perception be atypical in the three neurodevelopmental disorders? According to Guellaï, Streri, and Henny Yeung (2014; see also Streri, Coulon, Marie, & Henny Yeung, 2015), articulatory information (e.g., an individual's own silent articulations) influences AV speech perception. Because speech is delayed in DS, FXS, and WS, children with these neurodevelopmental disorders will experience less articulatory input, which will constrain the development of their AV speech perception skills and make it more difficult for them to acquire strong language skills. Furthermore, children with Down syndrome have anatomical and physiological differences in the mouth and throat that constrain development of oral motor control (Elliot & Weeks, 1993; Elliot, Weeks, & Gray, 1990; Kumin, 1996). Oral motor control is also atypical in FXS and WS (Barnes, Roberts, Mirrett, Sideris, & Misenheimer, 2006; Trauner, Bellugi, & Chase, 1989). Toddlers with WS often do go on to develop relatively proficient language skills, however. This may be reflected in our data, because whereas the CA-matched infants with WS discriminated between the congruent and incongruent faces in the Fusion condition (a pattern that was not observed in the TD controls), the mental and verbal age-matched toddlers with WS did not discriminate between the two faces in the Fusion condition and showed a more TD-like pattern of visual scanning. This hints at improvement in the WS developmental trajectory.

Why did we not obtain any significant results from the *fixation duration* analyses? We suspect that it is a consequence of participants experiencing lengthy video clips. The blocks of repetitions of mouth movements are indeed relatively long (12s). We speculate that the TD controls may have looked longer at one (e.g., the congruent) face for the first

several repetitions of mouth movements, and then—after several seconds—may have simply shifted the focus of their attention to the other (e.g., incongruent) face, resulting in equivalent looking times. Analyses of the first fixation showed this to be the case: TD controls were more likely to land a first fixation on the congruent face than on the incongruent face in the Mismatch condition, but not in the Fusion condition, as our hypothesis regarding AV integration would predict. In other words, the more time an infant has to scan a visual scene, the less sensitive becomes the measure of *fixation duration*, but an analysis of *first fixation* found that AV integration was present in the TD control group and yet demonstrably weak in all of the atypical groups. The reason that *fixation duration* yielded significant results for Guiraud et al. (2012) is arguably because Guiraud and colleagues tested younger children (8-9 month olds). Older children (such as the typically developing 15-16 month olds in the current study) are likely to require shorter presentation times to build an internal model of the familiar stimulus before switching attention to a novel stimulus. Therefore, a future experiment involving children older than 9-14 months may benefit from shorter presentation times and more trials.

In summary, our findings demonstrate that AV speech integration is weak in infants with WS, and that this might be the case also for the other two neurodevelopmental disorders but further research is required to verify this. Furthermore, whereas TD controls could discriminate between the faces in the Mismatch (but not Fusion) condition with their first fixation, the atypical groups all showed no discrimination in the Mismatch condition. Therefore, our results also hint that AV speech perception more generally (i.e., detection of incongruence between AV information) is poor across all atypically developing groups. Because AV speech matching and integration are key building blocks upon which higher-level skills (such as word learning) are built, we have identified a possible new constraint

on language development in WS (i.e., poor AV integration) and highlighted an area of future research (intermodal matching) that could be important for DS and FXS.

Acknowledgements

We are very grateful for the contributions families have made towards this study. The participation of the TD control participants in the research was supported by awards from the BASIS funding consortium led by Autistica (www.basisnetwork.org) and from the UK Medical Research Council (G0701484). The BASIS Team consists of (in alphabetical order): Tony Charman, Simon Baron-Cohen, Patrick Bolton, Kim Davies, Janice Fernandes, Jeanne Guiraud, Mark H. Johnson, Helen Maris, Helena Ribeiro and Leslie Tucker. The participation of participants with neurodevelopmental disorders was funded by the Waterloo Foundation, Williams Syndrome Foundation UK, Autour des Williams France, and a Wellcome Trust Strategic Award (grant number: 098330/Z/12/Z). The recruiting of participants was supported by the Williams Syndrome Foundation, the Fragile X Society, the Down Syndrome Association, and Down Syndrome Education International.

References

1. Alsius A, Navarra J, Campbell R, & Soto-Faraco S. (2005) Audiovisual integration of speech falters under high attention demands. *Curr Biol 15*: 839-843.
2. Amso, D., Haas, S., & Markant, J. (2014). An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes. *PLOS ONE*, *9*(1), e85701, 1-7.
3. Andersen, E. S., Dunlea, A., & Kekelis, L. S. (1984). Blind children's language: Resolving some differences. *Journal of Child Language*, *11*(3), 645-664.
4. Aslin, R. N. (2007). What's in a look?. *Developmental Science*, *10*(1), 48-53.

5. Bahrick, L. E. (1992). Infants' perceptual differentiation of amodal and modality-specific audio-visual relations. *Journal of Experimental Child Psychology*, *53*, 180-199.
6. Bahrick, L. E. (1994). The development of infants' sensitivity to arbitrary intermodal relations. *Ecological Psychology*, *6*, 111-123.
7. Bahrick, L. E. (2001). Increasing specificity in perceptual development: Infants' detection of nested levels of multimodal stimulation. *Journal of Experimental Child Psychology*, *79*, 253-270.
8. Bahrick, L. E. (2010). Intermodal perception and selective attention to intersensory redundancy: Implications for typical social development and autism. In G. Bremner & T. D. Wachs (Eds.), *The Wiley-Blackwell handbook of infant development, 2nd ed*. (pp. 120-166). Oxford, England: Blackwell Publishing.
9. Bahrick, L. E., Flom, R., & Lickliter, R. (2002). Intersensory redundancy facilitates discrimination of tempo in 3- month- old infants. *Developmental Psychobiology*, *41*(4), 352-363.
10. Bahrick, L. E. & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental Psychology*, *36*, 190-201.
11. Bahrick, L. E., & Lickliter, R. (2002). Intersensory redundancy guides early perceptual and cognitive development. *Advances in Child Development and Behavior*, *30*, 153-189.
12. Bahrick, L. E., & Todd, J. T. (2012). Multisensory processing in autism spectrum disorders: Intersensory processing disturbance as a basis for atypical development. In A. J. Bremner, D. J. Lewkowicz, & C. Spence (Eds.), *The new handbook of multisensory processes* (pp. 657-674). Oxford: Oxford University Press.

13. Basu Mallick, D., Magnotti, J. F., & Beauchamp, M. S. (2015). Variability and stability in the McGurk effect: contributions of participants, stimuli, time, and response type. *Psychonomic Bulletin Review*, 1-9. [Published online. DOI: 10.3758/s13423-015-0817-4.]

14. Böhning, M., Campbell, R., & Karmiloff-Smith, A. (2002). Audiovisual speech perception in Williams syndrome. *Neuropsychologia*, *40*(8), 1396-1406.

15. Bosch, L., & Sebastián-Gallés, N. (1997). Native-language recognition abilities in 4-month-old infants from monolingual and bilingual environments. *Cognition*, *65*(1), 33-69.

16. Burnham, D., & Dodd, B. (2004). Auditory-visual speech integration by prelinguistic infants: Perception of an emergent consonant in the McGurk effect. *Developmental Psychobiology*, *45*(4), 204-220.

17. Burr, D., & Alais, D. (2006). Combining visual and auditory information. *Progress in Brain Research*, *155*, 243-258.

18. Campbell, R. (1996, October). Seeing speech in space and time: Psychological and neurological findings. In *Spoken Language, 1996. ICSLP 96. Proceedings., Fourth International Conference on* (Vol. 3, pp. 1493-1496). IEEE.

19. Campbell, R. (2008). The processing of audio-visual speech: empirical and neural bases. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *363*(1493), 1001-1010.

20. Chien, S. H-L. (2011). No more top-heavy bias: Infants and adults prefer upright faces but not top-heavy geometric or face-like patterns. *Journal of Vision*, *11*(6), 13.

21. Cohen, L. B. (1972). Attention-getting and attention-holding processes of infant visual preferences. *Child Development*, *43*, 869-879.

22. Crumrine, D., Owens, J., Adams, M., & Salamone, L. (2010). A preliminary investigation of eye gaze patterns on fast-mapping abilities of children with ASD. In *Proceedings of the 6th Annual GRASP Symposium* (pp. 93-94). Wichita, KS: Wichita State University.
23. Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants?. *Developmental psychobiology*, *45*(4), 187-203.
24. Dodd, B. (1979). Lip reading in infants: Attention to speech presented in-and out-of-synchrony. *Cognitive Psychology*, *11*(4), 478-484.
25. D'Souza, D., D'Souza, H., Johnson, M. H., & Karmiloff-Smith, A. (2015). Concurrent relations between face scanning and language: A cross-syndrome infant study. *PLOS ONE*, *10*(10), e0139319.
26. Fenson, L., Dale, P. S., Reznick, J. S., Thal, D., Bates, E., Hartung, J. P., … & Reilly, J. S. (1993). *The MacArthur Communicative Development Inventories: User's guide and technical manual*. San Diego, CA: Singular Publishing Group.
27. Field, A. (2013). *Discovering statistics using IBM SPSS statistics*. London, UK: Sage.
28. Gliga, T., Elsabbagh, M., Andravizou, A., & Johnson, M. (2009). Faces attract infants' attention in complex displays. *Infancy*, *14*(5), 550-562.
29. Golinkoff, R. M., Hirsh-Pasek, K., Cauley, K. M., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, *14*, 23-45.
30. Guastella, A. J., Mitchell, P. B., & Dadds, M. R. (2008). Oxytocin increases gaze to the eye region of human faces. *Biological Psychiatry*, *63*, 3-5.

31. Guiraud, J. A, Tomalski, P., Kushnerenko, E., Ribeiro, H., Davies, K., Charman, T., … & Johnson, M. H. (2012). Atypical audiovisual speech integration in infants at risk for autism. *PloS One*, *7*(5), e36428.

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

33. Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, *2*(10), 389-398.

34. Karmiloff-Smith, A., D'Souza, D., Dekker, T. M., Van Herwegen, J., Xu, F., Rodic, M., & Ansari, D. (2012). Genetic and environmental vulnerabilities in children with neurodevelopmental disorders. *Proceedings of the National Academy of Sciences*, *109*(Supplement 2), 17261-17265.

35. Vorperian, H. K., & Kent, R. D. (2007). Vowel acoustic space development in children: A synthesis of acoustic and anatomic data. *Journal of Speech, Language, and Hearing Research*, *50*(6), 1510-1545.

36. Kuhl, P. K., & Meltzoff, A. N. (1982). The bimodal perception of speech in infancy. *Science*, *218*(4577), 1138-1141.

37. Kuhl, P. K., & Meltzoff, A. N. (1984). The intermodal representation of speech in infants. *Infant Behavior and Development*, *7*(3), 361-381.

38. Kuhl, P. K., & Meltzoff, A. N. (1988). Speech as an intermodal object of perception. In A. Yonas (Ed.), *Perceptual development in infancy: The Minnesota symposia on child psychology* (Vol. 20, pp. 235-266). Hillsdale, NJ: Lawrence Erlbaum.

39. Kuhl, P. K., Williams, K. A., & Meltzoff, A. N. (1991). Cross-modal speech perception in adults and infants using nonspeech auditory stimuli. *Journal of Experimental Psychology: Human Perception and Performance*, *17*(3), 829.

40. Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences of the United States of America, 105*(32), 32-35.

41. Lewkowicz, D. J. (2000). The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psychological Bulletin*, *126*(2), 281.

42. Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental Psychology*, *46*(1), 66.

43. Lewkowicz, D. J., Leo, I., & Simion, F. (2010). Intersensory perception at birth: Newborns match nonhuman primate faces and voices. *Infancy*, *15*(1), 46-60.

44. MacKain, K., Studdert-Kennedy, M., Spieker, S., & Stern, D. (1983). Infant intermodal speech perception is a left-hemisphere function. *Science*, *219*(4590), 1347-1349.

45. Mani, N., Mills, D. L., & Plunkett, K. (2012), Vowels in early words: an event-related potential study. *Developmental Science*, *15*, 2–11.

46. Marcell, M. M., & Armstrong, V. (1982). Auditory and visual sequential memory of Down syndrome and nonretarded children. *American Journal of Mental Deficiency*, *87*(1), 86-95.

47. Massaro, D. W. (1987). *Speech perception by ear and eye: a paradigm for psychological enquiry*. Hillsdale NJ: Lawrence Erlbaum.

48. Massaro, D. W. (1998). *Perceiving talking faces: from speech perception to a behavioural principle*. Cambridge, MA: MIT Press.

49. Maurer, D., & Salapatek, P. (1976). Developmental changes in the scanning of faces by young infants. *Child Development*, *47*, 523–527.

50. McCune, B., & Grace, J. B. (2002). *Analysis of ecological communities*. Gleneden Beach, OR: MjM Software.

51. McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746-748.

52. Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertoncini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, *29*(2), 143-178.

53. Mertens, I., Siegmund, H., &, Grusser, O. J. (1993). Gaze motor asymmetries in the perception of faces during a memory task. *Neuropsychologia*, *31*(9), 989-998.

54. Morrongiello, B. A., Fenwick, K. D., & Chance, G. (1998). Crossmodal learning in newborn infants: Inferences about properties of auditory-visual events. *Infant Behavior and Development*, *21*(4), 543-553.

55. Mullen, E. (1995). *Mullen Scales of Early Learning* (AGS ed.). Circle Pines, MN: American Guidance Service.

56. Nardini, M., Bedford, R., & Mareschal, D. (2010). Fusion of visual cues is not mandatory in children. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(39), 17041-17046.

57. Noton, D., & Stark, L. (1971). Scan paths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, *11*, 929–942.

58. Paterson, S. J., Brown, J. H., Gsodl, M. K., Johnson, M. H., & Karmiloff-Smith, A. (1999). Cognitive modularity and genetic disorders. *Science*, *286*(5448), 2355-2358.

59. Patterson, M. L., & Werker, J. F. (1999). Matching phonetic information in lips and voice is robust in 4.5-month-old infants. *Infant Behavior and Development*,*22*(2), 237-247.

60. Patterson, M. L., & Werker, J. F. (2003). Two-month-old infants match phonetic information in lips and voice. *Developmental Science*, *6*(2), 191-196.

61. Pérez-Pereira, M., & Conti-Ramsden, G. (2013). *Language development and social interaction in blind children*. Hove, UK: Psychology Press.

62. Pons, F., Lewkowicz, D. J., Soto-Faraco, S., & Sebastián-Gallés, N. (2009). Narrowing of intersensory speech perception in infancy. *Proceedings of the National Academy of Sciences*, *106*(26), 10598-10602.

63. Poole, A., Ball, L. J., & Phillips, P. (2004). In search of salience: A response time and eye movement analysis of bookmark recognition. In S. Fincher, P. Markopolous, D. Moore, & R. Ruddle (Eds.), *People and computers XVIII-design for life: Proceedings of HCI 2004*. London: Springer-Verlag.

64. Porter, M. A., Shaw, T. A., & Marsh, P. J. (2010). An unusual attraction to the eyes in Williams-Beuren syndrome: a manipulation of facial affect while measuring face scanpaths. *Cognitive Neuropsychiatry*, *15*(6), 505-530.

65. Riby, D. M., & Hancock, P. J. B. (2008). Viewing it differently: Social scene perception in Williams syndrome and autism. *Neuropsychologia*, *46*(11), 2855-2860.

66. Riby, D. M., & Hancock, P. J. B. (2009a). Do faces capture the attention of individuals with Williams syndrome or Autism? Evidence from tracking eye movements. *Journal of Autism and Developmental Disorders*, *39*(3), 421-431.

67. Riby, D. M., & Hancock, P. J. B. (2009b). Looking at movies and cartoons: Eye-tracking evidence from Williams syndrome and autism. *Journal of Intellectual Disability Research*, *53*(2), 169-181.

68. Rice, M. L., Warren, S. F., & Betz, S. K. (2005). Language symptoms of developmental language disorders: An overview of autism, Down syndrome, fragile

X, specific language impairment, and Williams syndrome. *Applied Psycholinguistics*, *26*(1), 7-27.

69. Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & Psychophysics*, *59*(3), 347-357.
70. Rupp, H. A., & Wallen, K. (2007). Sex differences in viewing sexual stimuli: An eye-tracking study in men and women. *Hormones and Behavior*, *51*, 524-533.
71. Scerif, G., Longhi, E., Cole, V., Karmiloff-Smith, A., & Cornish, K. (2012). Attention across modalities as a longitudinal predictor of early outcomes: the case of fragile X syndrome. *Journal of Child Psychology and Psychiatry*, *53*(6), 641-650.
72. Schiller, P. H., & Tehovnik, E. J. (2001). Look and see: How the brain moves your eyes about. *Progress in Brain Research*, *154*, 127-142.
73. Schroeder, C. E., Lakatos, P., Kajikawa, Y., Partan, S., & Puce, A. (2008). Neuronal oscillations and visual amplification of speech. *Trends in Cognitive Science*, *12*(3), 106-113.
74. Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the* Acoustical *Society of America*, *26*(2), 212-215.
75. Teinonen, T., Aslin, R. N., Alku, P., & Csibra, G. (2008). Visual speech contributes to phonetic learning in 6-month-old infants. *Cognition*, *108*(3), 850-5.
76. Tomalski, P., Ribeiro, H., Ballieux, H., Axelsson, E. L., Murphy, E., Moore, D. G., & Kushnerenko, E. (2013). Exploring early developmental changes in face scanning patterns during the perception of audiovisual mismatch of speech cues. *European Journal of Developmental Psychology*, *10*(5), 611–624.
77. Tuomainen, J., Andersen, T. S., Tiippana, K., & Sams, M. (2005). Audio-visual speech perception is special. *Cognition*, *96*, B13-22.

78. van Atteveldt, N., Murray, M. M., Thut, G., & Schroeder, C. E. (2014). Multisensory integration: Flexible use of general operations. *Neuron*, *81*, 1240-1253.

79. Van der Molen, M. J. W., Van der Molen, M. W., Ridderinkhof, K. R., Hamel, B. C. J., Curfs, L. M. G., & Ramakers, G. J. A. (2012). Auditory and visual cortical activity during selective attention in fragile X syndrome: a cascade of processing deficiencies. *Clinical Neurophysiology*, *123*(4), 720-729.

80. Vroomen, J., & Stekelenburg, J. J. (2011). Perception of intersensory synchrony in audiovisual speech: Not that special. *Cognition*, *118*(1), 75-83.

81. Walker-Smith, G. J., Gale, A. G., & Findlay, J. M. (1977). Eye movement strategies in face perception. *Perception*, *6*, 313–326.

82. Wallace, M. T., & Stevenson, R. A. (2014). The construct of the multisensory temporal binding window and its dysregulation in developmental disabilities. *Neuropsychologia*, *64*, 105-123.

83. Walton, G. E., & Bower, T. G. R. (1993). Newborns form "prototypes" in less than 1 minute. *Psychological Science*, *4*(3), 203-205.

84. Weikum, W. M., Vouloumanos, A., Navarra, J., Soto-Faraco, S., Sebastián-Gallés, N., & Werker, J. F. (2007). Visual language discrimination in infancy. *Science*, *316*(5828), 1159.

85. Wilson, C. E., Palermo, R., & Brock, J. (2012). Visual scan paths and recognition of facial identity in autism spectrum disorder and typical development. *PloS One*, *7*(5), e37681.

86. Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychological Science*, *24*(5), 603-612.