Infant Behavior and Development 60 (2020) 101448

Contents lists available at ScienceDirect

Infant Behavior and Development

journal homepage: www.elsevier.com/locate/inbede

Contrasting behavioral looking procedures: a case study on infant speech segmentation

Caroline Junge[a,*], Emma Everaert[b], Lyan Porto[c], Paula Fikkert[c], Maartje de Klerk[b], Brigitta Keij[b], Titia Benders[d]

[a] Departments of Experimental and Developmental Psychology, Utrecht University, Utrecht, The Netherlands
[b] Utrecht Institute of Linguistics OTS, Utrecht University, Utrecht, The Netherlands
[c] Centre for Language Studies, Radboud University, Nijmegen, The Netherlands
[d] Department of Linguistics, Macquarie University, Sydney, Australia

ARTICLE INFO

Keywords:
Infant preference
Central fixation
Headturn preference procedure
Speech segmentation ability
Familiarity response

ABSTRACT

This paper compared three different procedures common in infant speech perception research: a headturn preference procedure (HPP) and a central-fixation (CF) procedure with either automated eye-tracking (CF-ET) or manual coding (CF-M). In theory, such procedures all measure the same underlying speech perception and learning mechanisms and the choice between them should ideally be irrelevant in unveiling infant preference. However, the ManyBabies study (ManyBabies Consortium, 2019), a cross-laboratory collaboration on infants' preference for child-directed speech, revealed that choice of procedure can modulate effect sizes. Here we examined whether procedure also modulates preference in paradigms that add a learning phase prior to test: a speech segmentation paradigm. Such paradigms are particularly important for studying the learning mechanisms infants can employ for language acquisition. We carried out the same familiarization-then-test experiment with the three different procedures (32 unique infants per procedure). Procedures were compared on various factors, such as overall effect, average looking time and drop-out rate. The key observations are that the HPP yielded a larger familiarity preference, but also reported larger drop-out rates. This raises questions about the generalizability of results. We argue that more collaborative research into different procedures in infant preference experiments is required in order to interpret the variation in infant preferences more accurately.

1. Introduction

Children's speech perception undergoes marked development throughout infancy (Kuhl et al., 2008; Werker & Hensch, 2015). Much of what we have learned about infant speech perception has come from behavioral procedures: that is, from measuring infants' reactions while they listen to different types of speech. Although it is the audio that chang es from trial to trial during test, researchers often rely on infants' looking behavior to an unrelated visual stimulus as a proxy of interest in the accompanying speech fragment. It is inferred that infants are able to discriminate between two types of speech stimuli when they respond differently to these stimuli, which typically means a preferential response to one type of speech stimulus over the other. However, an absence of preference does not need to imply an absence of discrimination ability (Aslin, 2007).

* Corresponding author.
E-mail address: C.M.M.Junge@uu.nl (C. Junge).

https://doi.org/10.1016/j.infbeh.2020.101448
Received 31 March 2018; Received in revised form 21 December 2019; Accepted 3 April 2020

Research shows that most infants show surprisingly consistent preferences without training in the lab for some experimental comparisons. For instance, infants prefer to hear infant-directed speech over adult-directed speech (Cooper & Aslin, 1990; ManyBabies Consortium, 2019); to hear their native language over an unfamiliar language (Moon, Cooper, & Fifer, 1993; Nazzi, Jusczyk, & Johnson, 2000); and to hear familiar words over unknown words (Shi, Werker, & Cutler, 2006; Swingley, 2005). Such studies reveal what children have learned through prior exposure to their native language(s). Other studies add a learning phase prior to test, exposing infants to particular phonemes (e.g., Maye, Werker, & Gerken, 2002; Yoshida, Pons, Maye, & Werker, 2010) or word forms (Jusczyk and Aslin, 1995) before testing whether infants can discriminate between stimuli that either reappeared from the learning phase ('old') or not ('novel'). These familiarization-then-test procedures enable researchers to tap into the learning mechanisms and biases that children employ in learning language.

There are a number of existing testing procedures that rely on infants' looking preference, which can differ in the type of looking response that is elicited (i.e., headturns vs. eye movements) as well as the method used for recording these responses (i.e., manually-controlled vs. automatic). In theory, such procedures all measure the same underlying speech perception and learning mechanisms; consequently, the choice between them should ideally be irrelevant in unveiling infant preferences for one speech type over another. Yet, whether the choice of method affects the presence, strength, or direction of infant preferences is still an empirical question. Answering this question is important for interpreting not only the outcomes of individual studies, but also for understanding differences across studies in systematic literature reviews or meta-analyses. In the following, we first describe the relevant testing procedures in more detail, before explaining how the present study sets out to directly compare these procedures using the same speech stimuli and design.

One of the most widely used testing procedures to index infant looking behavior is the headturn preference procedure (HPP; Fernald, 1985). The HPP works broadly as follows: a child is sitting on her parent's lap in the center of a three-sided booth outfitted with one center light on the wall facing the participant and two lights on the side walls. When the infant orients to the flashing center light for a set minimum amount of time, this center light is extinguished and one of the side lights starts flashing. When the infant turns her head to orient to the flashing side light, an auditory stimulus starts playing. The ending of the auditory stimulus generally coincides with that of the visual stimulus: both stop when the child looks away for a set amount of time (e.g., 2 seconds), or when the maximum trial duration has been reached (e.g., after 20 seconds). Thus, presentation of speech stimuli is usually infant-controlled as it is contingent on how long the infant turns her head to attend to a flashing light. Infant looking behavior in the HPP is quantified as the amount of time the infant has spent with a turned head, looking at the flashing side light, while the auditory stimulus was playing.

A second means to assess infant looking behavior is the central fixation procedure (CF; Cooper & Aslin, 1990). In this procedure the infant faces a single screen depicting an unrelated visual stimulus, such as a dynamic checkerboard, while simultaneously speech stimuli are played (e.g., Stager & Werker, 1997). Usually, a different visual stimulus, an 'attention getter' is presented between trials to clearly distinguish the trials from one another. Infant looking behavior in the CF is the amount of time during each trial that the child looks at the visual stimulus. Trial endings can be infant-controlled, as in the HPP, but some CF studies play auditory stimuli until the maximum trial duration has been reached, even when the child is inattentive to the screen (e.g., Best & Jones, 1998; ter Schure, Junge, & Boersma, 2016).

The HPP and the CF procedure both rely on the same principle: infant looking preferences are indicative of auditory preferences. Yet, these procedures differ in two key aspects of their quantification of visual interest. First, the HPP focuses on *gross body* changes whereas the CF procedure also considers *fine ocular* movements. Second, the HPP begins to calculate looking time when the child is looking *away* from the forward-facing position, whereas the CF procedure requires the child to keep looking *towards* a visual display in the forward-facing position.

In addition, one can implement the CF procedure to record infants' looking behavior in one of two ways: either monitoring the child's eyes through a peephole or closed-circuit video for online coding and videotaping the child's eyes for subsequent offline coding by human observers ('CF-manual procedure') or using an automated eye tracker that tracks the child's gaze direction in real-time ('CF-eye tracking procedure'). Each implementation comes with its own advantages and drawbacks. The field of infant research is increasingly using automatic eye trackers in tests of infant cognition, as it is considered less time-consuming and more reliable than human coding (Aslin, 2012; Gredebäck, Johnson, & von Hofsten, 2009; Oakes, 2012). However, potential downsides of automatic eye trackers for testing infants have recently been acknowledged as well (Hessels & Hooge, 2019). One concern is that eye trackers may have difficulty tracking infants' eyes when infants tilt their heads from the start position, making data less accurate than advertised by publishers and possibly resulting in data loss even when the children are fixating the screen (Hessels, Andersson, Hooge, Nyström, & Kemner, 2015). Such data loss would be regarded as a loss of visual interest in the CF procedure, possibly impacting on the accuracy of quantifying infants' preference for speech types. However, it is currently unclear whether there is a difference between manual and automatic CF procedures in their ability to detect the presence and strength of infant preferences for one type of speech over another.

There is reason to believe that this variation in procedures contributes to the variation that is typically observed in infant studies (ManyBabies Consortium, 2019). The ManyBabies Consortium recently conducted a large-scale multi-lab study aimed at better understanding how a wide range of factors, including differences between testing procedures, contributes to the observed variation in outcomes of infant studies. The main effect of interest was infants' preference for infant-directed speech over adult-directed speech, which is a well-established and robust effect in the literature (Dunst, Gorman, & Hamby, 2012). Crucially, the 67 participating labs all tested this preference using the same auditory stimuli, but typically used only one of the three procedures described above to index infant preference. While the manual and automatic CF procedure yielded similar infant preferences – each eliciting small but significant preferences for infant-directed speech –, this preference was significantly stronger for labs using the HPP. These results suggest that the contrast between the HPP and CF procedure impacts on the strength of the detected infant preference, but that the contrast between the manual and automatic implementation of the CF procedure may be less critical.

Infants' preference for infant-directed over adult-directed speech is presumably the result of infants' long-term exposure to speech – knowledge that they take to the lab visit. Therefore, the test phase to assess this preference can commence (almost) immediately after the start of the procedure without the need for a familiarization phase. The current study asks whether the choice of procedure also affects infant preference for stimuli that *have* versus *have not* been introduced in a familiarization phase preceding the test phase. The effect of the testing procedure might be less pronounced or absent for familiarization-then-test designs, as the familiarization phase typically offers infants an opportunity to become acquainted with the procedure, including the contingency between their looking behavior and the presentation of the auditory stimulus. Yet, if the presence of a familiarization phase does impact on the strength of the detected effect, the prediction would be that the HPP elicits a larger preference than both the manual and the automatic CF procedure. To directly address this issue, we carried out the same experimental design, with three different testing procedures involving three different yet comparable sets of 32 10-month-olds: the HPP (Experiment 1), the CF-eye tracking procedure (CF-ET, Experiment 2) and the CF-manual procedure (CF-M, Experiment 3).

2. The current study: rationale and goals

The familiarization-then-test design has been very frequently employed to test infant speech segmentation skill, following Jusczyk and Aslin's (1995) first demonstration of both the suitability of the paradigm and the presence of speech segmentation skills in infants. In such speech segmentation tests, either one of the phases (familiarization or test phase) comprises passages with target words embedded in speech, while the other consists of repetitions of single words. The test phase always represents words from the familiarization phase ('familiar' words) as well as phonologically matched words not presented in the familiarization phase ('novel' words). A differential looking time to familiar and novel words at test indicates that infants have recognized the familiar words across the phases, and hence have segmented the passages into separate words. Note, however, that it is more difficult to predict the direction of infants' preference in speech segmentation studies than in studies contrasting responses to infant-directed versus adult-directed speech (Frank et al., 2017). The speech segmentation literature documents preferences for the familiar words (familiarity preference), the novel words (novelty preference) (cf. Bergmann & Cristia, 2016), as well as null effects (Floccia et al., 2016).

One way to address the impact of testing procedures on outcomes is via meta-analyses. There is one recently published meta-analysis on possible sources of variation in infant speech segmentation studies with natural speech, which focused on age and task difficulty as possible moderators, but did not examine type of procedure as a possible factor (Bergmann & Cristia, 2016). Fortunately, this meta-analysis is available online through Metalab, which is continuously updated with new studies (available online: https://metalab.stanford.edu; cf. Bergmann et al., 2018). Note that this database lists the response mode (eye tracking vs. headturns) but does not differentiate between manual or automatic eye tracking.

To understand whether the testing procedure has an impact on group performance, we first analyzed all existing word segmentation studies (n = 274) with native speech stimuli as stored in Metalab (Bergmann et al., 2018), accessed on December 19, 2019. Following Bergmann and Cristia (2016), we focused on Hedges' g. There is sufficient variation in the data to warrant a search for additional moderators (Q (273) = 1071.1, p < .001). Adding infant response mode (eye tracking vs. headturns) as a possible mediator improves the model, albeit not convincingly (QM (1) = 1.40, p = 0.24). Effect sizes for eye tracking studies were numerically smaller than effect sizes obtained in HPP studies, as indicated by the negative estimate for the eye tracking method (β = -0.088; SE = 0.08) compared to the HPP. Nonetheless, both types of testing procedures yield average positive effect sizes (HPP: Hedges' g =0.251; SE = .04; eye tracking: Hedges' g = 0.129; SE = .03). In interpreting the non-significant effect of procedure on variation in infant preferences, it is important to note that substantially fewer records measured eye tracking (n = 37) compared to headturns (n = 237). Hence, it could be that with more studies using eye tracking as a means to tap into infant segmentation ability it might become clear that type of procedure in fact moderates infant preferences in learning paradigms. Moreover, experiments differed in choice of auditory speech stimulus, task difficulty, and age of children. A direct comparison between procedures using the same auditory stimuli, as well as the same population and experimental parameters, is needed to shed more light on the impact of the different behavioral procedures on the strength of the detected preferences in a familiarization-then-test speech segmentation task, which is what we set out to do in the current paper.

Our design closely followed Jusczyk, Houston, and Newsome, Experiment 2 (1999), who used the HPP to test American-English 7-to-8-month-olds on their ability to recognize low-frequency trochaic words after being exposed to these words in passages during familiarization. We opted for presenting passages in the familiarization and single words at test (rather than the reverse order) because some studies suggest that this order may yield larger effects (e.g., Nazzi, Mersad, Sundara, Iakimova, & Polka, 2014; van Heugten & Johnson, 2012). We implemented three modifications compared to the Jusczyk, Houston, and Newsome (1999) study. First, our target words were pseudowords that followed the native language's phonotactics instead of low-frequency words, to ensure that each target would be truly unknown to infants prior to the experimental session. Second, since we tested Dutch infants, we created Dutch stimuli recorded in exaggerated infant-directed speech (Floccia et al., 2016; Schreiner & Mani, 2017). Third, we tested infants of 10 months of age, because Dutch infants are slightly delayed in their segmentation abilities, compared to their American-English peers (Houston, Jusczyk, Kuijpers, Coolen, & Cutler, 2000), but are able to segment disyllabic words with the trochaic stress pattern typical for Dutch by 10 months of age (Kooijman, Hagoort, & Cutler, 2009).

The HPP version of this experiment was created before implementing the two CF procedures (one with automated eye tracker; the other with manual operators), each of which resembles the HPP version as closely as possible bar its dependent measure (e.g., headturns vs. visual fixations). While the HPP was conducted at one university, the two CF procedures were conducted at another university. Both testing facilities were dedicated to infant research ('babylabs'), located near or in the center of university towns

separated at 68 kilometers from each other. Both babylabs have ample experience with testing infants in the respective procedure that they executed for this comparative study.

It was hypothesized that infants would show a familiarity preference across all three procedures. We hypothesized that the strength of detected infant preferences in a familiarization-then-test segmentation procedure might hinge on the type of testing procedure employed, with stronger preferences being detected in the HPP (ManyBabies Consortium, 2019). However, it was also conceivable that the testing procedure would not moderate preferences that are invoked in the lab, in which case the strength of the detected preference would be similar across the three procedures. Finally, besides directly comparing the three procedures on infant preferences for the test stimuli, we also explored whether they differ on other variables that may reflect participants' performance in the procedure, such as the participant dropout rates, the number of familiarization trials until participants reached the familiarization criterion, participants' overall duration of attention to test trials, and the number of test trials that participants completed successfully.

3. Methods

3.1. Experiment 1: Headturn Preference Procedure (HPP)

3.1.1. Subjects

A total of 32 full-term, typically developing (i.e., without a family history of dyslexia) Dutch monolingual 10-month-olds (16 girls; M_{age} = 303 days, range 286 – 318 days) contributed sufficient data to this experiment. Another 36 infants were tested but excluded due to the following reasons: inattentive to lights (n = 17); started crying (n = 12); technical problems (n = 4); standing on parent's lap so infant headturns were out of the observer's vision (n = 2); or parental interference (n = 1). Participants were recruited via the babylab database of the Radboud University Nijmegen. Ethical approval for the study was obtained from the local Ethics Committee. All parents provided written informed consent for their child to participate and they received a book or 10 euro in appreciation of their participation.

3.1.2. Stimuli

Two trochaic pseudowords that followed Dutch phonotactics from another word-learning experiment in Dutch served as the basis of our target word pairs (/taːsəl/ and /tɑnoː/; Tsuji, Fikkert, Yamane, & Mazuka, 2016). We created four different versions of this target pair by replacing the word-initial consonants[1] . Word onsets of each target pair shared place of articulation (both labial or both alveolar) but differed in voicing (voiced versus voiceless). This resulted in four target pairs, which were presented between subjects: 1. /paːsəl/-/bɑnoː/; 2. /baːsəl/-/pɑnoː/; 3. /daːsəl/-/tɑnoː/; and 4. /taːsəl/-/dɑnoː/ as familiar words. As novel words we selected two trochaic pseudowords that were created to be segmentally distinct from the target words (i.e., segmental content of the stressed syllables did not overlap between target and novel words): /xeːmər/ and /foːni/.

Two sets of six-sentence passages were created to be recorded with each of the eight target words (see Table 1). The target words were preceded by the same six words in both passages. The position of the target word in the carrier sentences was balanced across the two passages: The target word was presented near the beginning of two sentences, in the middle of two sentences, and in final position of two sentences. The target word appeared in each sentence position in the first three as well as in the last three sentences of each passage.

One female native Dutch speaker produced all sentences and single words in an enthusiastic infant-directed speech register. All stimuli were recorded in mono at 44.1 kHz in a sound-proof chamber and saved as WAV files (705kbps). Praat (Boersma & Weenink, 2015) was used to create passages and word lists. Mean length of passage A is 32.72 seconds (range 32.09 – 33.27) and mean length of passage B is 33.61 seconds (range 29.64 – 37.21). Word lists consist of sixteen instances of the word in isolation (8 tokens). Each token appears twice in the list in pseudorandom order albeit never in direct repetition. The inter-stimulus interval between the tokens is approximately 1 second and the lists always start and end with a 500-millisecond silence. These word lists are on average 35.53 seconds long (range 32.61 – 37.52). All recordings are equalised to the same mean level of intensity, i.e., loudness (69 dB).

3.1.3. Design

There were 16 versions of the experiment, counterbalancing Word Pairs as targets (4) x Passages (2; -aːsəl embedded in passage A or passage B) x Order of target in familiarization (2; -aːsəl in presented first or second in the familiarization; see Table 2). Each version was presented to at least one infant, and to no more than three infants. All trials were infant-controlled, playing to completion as long as a child maintained interest in a light or stopping once a child looked away for more than 2 consecutive seconds. The familiarization phase comprised a maximum of 18 trials presenting passages with the target words in alternating order. Familiarization stopped when at the end of a trial a child had been exposed to each word for at least 45 seconds. If the criterion of 45 seconds was reached for one word, but not yet for the other, familiarization continued with only the passage containing the word for which the familiarization criterion had not yet been reached. Infants were exposed to each target word at least nine times during the familiarization, before moving on to the test phase.

The test phase consisted of three blocks of four trials. Each test trial played a word list with either one of the two familiar words or one of the two novel words. All words appeared once within each block. Trials of a certain word were never played consecutively

[1] This was done because this experiment would additionally serve as a baseline for other experiments examining consonant contrasts.

Table 1
The two passages used in the familiarization phase. Target word position is indicated by […]; Words preceding targets are underlined.

Passage A	*Translation*
Jouw […] ligt al boven op zolder.	Your […] is already lying in the attic.
De boer gebruikte vaak die […] als hij op bezoek kwam.	The farmer often used this […] when he came to visit.
Eens zag hij een mier bij die oude […].	Once he saw an ant near the old […].
De […] kreeg al snel meer gaten.	The […] quickly got more holes.
Dus wil hij een andere […].	So he wants another […].
Met de nieuwe […] is hij blijer.	With the new […] he is happier.
Passage B	*Translation*
Jouw […] is erg geweldig.	Your […] is really amazing.
Iemand anders zag laatst een mooie, oude […].	Someone else recently saw a pretty, old […].
Mensen zien de […] dan niet eens.	People just do not even see the […].
Een andere […] is vaak bruiner.	Another […] is often browner.
Mensen met die […] gaan graag naar buiten.	People with that […] love to go outside.
Vaak worden ze dan gezien als een nieuwe […].	Often they are then perceived as a new […].

Table 2
The properties of the familiarization phase of the sixteen versions of the experiment.

Passage	word pair 1: /baːsəl/-/pɑnoː/	word pair 2: /paːsəl/-/bɑnoː/	word pair 3: /taːsəl/-/dɑnoː/	word pair 4: /daːsəl/-/tɑnoː/
-aːsəl- in *A*, -ɑnoː in *B*	baːsəl, pɑnoː	paːsəl, bɑnoː	taːsəl, dɑnoː	daːsəl, tɑnoː
-aːsəl- in *B*, -ɑnoː in *A*	baːsəl, pɑnoː	paːsəl, bɑnoː	taːsəl, dɑnoː	daːsəl, tɑnoː
-aːsəl- in *A*, -ɑnoː in *B*	pɑnoː, baːsəl	bɑnoː, paːsəl	dɑnoː, taːsəl	tɑnoː, daːsəl
-aːsəl- in *B*, -ɑnoː in *A*	pɑnoː, baːsəl	bɑnoː, paːsəl	dɑnoː, taːsəl	tɑnoː, daːsəl

Note: The target word in the left of each cell was presented first during familiarization, after which the target words were alternated.

(e.g., at the last trial of block 1 and the first trial of block 2). The order of trials within a block was never repeated in the next block, and the ordering of trials within each block was pseudorandom and counterbalanced across the different versions. The test phase ended when the infant reached the end of the twelfth trial.

3.1.4. Apparatus

Testing took place in a three-sided white pegboard booth in a dimly lit room. The wall directly in front of the child was outfitted with a blue light and the walls on either side were outfitted with red lights. A camera (Mini CCTV camera MC 900-D12) peeping through a 5 cm diameter hole below the center light was used to monitor and record the child's headturns. These video recordings were stored on a computer (HP Compaq dc7700) to allow for off-line coding. Two speakers (Canton LE-101) below the side lights played the audio stimuli. The sound files were stored on another computer (HP Compaq dc7700) and played through an amplifier (Sony TA-FE230) configured to output sound between 65 and 70 dBA, as measured by a decibel meter. The experimenter stood behind the curtain, wearing headphones playing masking noise, so that he or she was blind to the stimuli being played. The experimenter monitored a live feed of the child on a television set (Sony Trinitron) and pressed buttons on the computer keyboard corresponding to the direction in which the infant was looking. The button presses were recorded by the software LOOK (Meints & Woodford, 2008), which also controlled the presentation of audio stimuli and appropriate illumination of the different lamps and generated experiment log files, including data on participants' headturns. This process, called 'on-line coding,' served to control that stimuli were only played when the child's gaze was directed to the flashing light, which is taken to represent the child's attention.

3.1.5. Procedure

Upon arrival, the experiment was explained to parents, who then provided informed consent. The child sat on a parent's lap on a chair in the center of the test booth. The parent was instructed not to talk, point or look at the child or the side lights, but to keep his or her gaze directed forwards for the entire duration of the experiment while wearing headphones (Sennheiser HMEC 300) with masking speech so that they were blind to the stimuli being played and could not interfere with the experiment.

All trials began with the center light flashing to grab the infant's attention. Once the infant oriented to that light for 1 uninterrupted second, the light extinguished, and one of the two side lights began to flash. When the infant turned her head at least 30° to the side of the flashing lamp, and the experimenter pressed the corresponding button, the audio stimulus started playing from the speaker below the flashing light. If the child turned her head away from the flashing light for more than 2 seconds, or if the stimulus had been played to completion, the trial ended and the center light began flashing again in preparation for the next trial. The sides on which the stimuli were played were pseudorandomized and counterbalanced across trials.

The experiment began with the familiarization phase. The test phase started immediately after the end of the familiarization

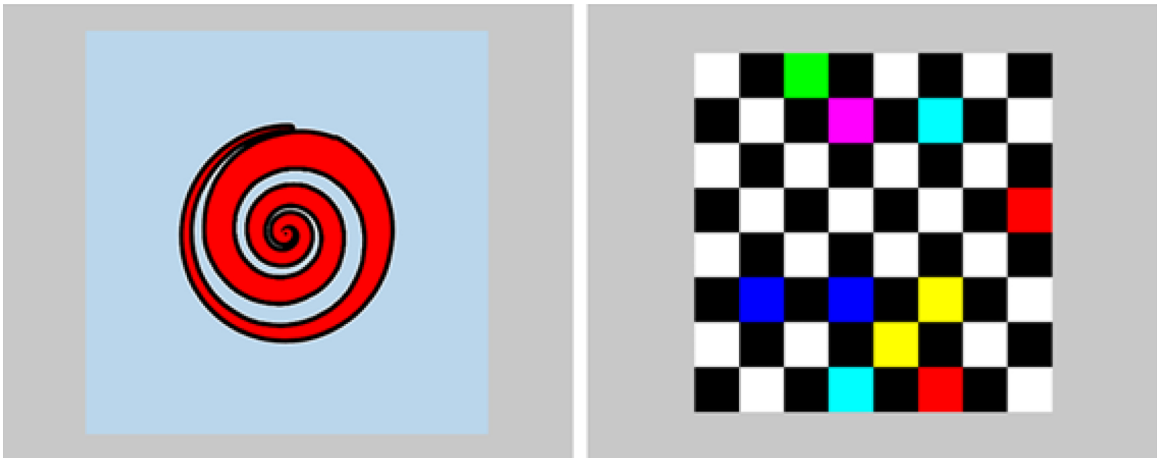

Fig. 1. The visual stimuli used in both CF experiments. Left: The attention getter. Right: One of the 63 checkerboard images visible during the presentation of the auditory stimuli.

phase. The total duration of the experiment depended on the child's behaviour but usually fell between 5 to 7 minutes.

3.1.6. Pre-processing steps

We used ELAN (Sloetjes & Wittenburg, 2008) for off-line coding of infants' headturns to the flashing lights. Coders were blind to the type of words infants heard. Cases of ambiguous behaviour were discussed until researchers reached agreement. Looks away from the flashing lights were discarded from the total looking time, also when the look away had been too short (<2 seconds in online coding) to terminate the trial. Trials with a total looking time of less than 1 second were removed. Infants were only retained for further analyses if they contributed at least three test trials per condition (familiar/novel words).

3.2. Experiment 2: Central Fixation procedure using automated eye tracking (CF-ET)

3.2.1. Subjects

All 32 infants who contributed data were full-term, typically developing Dutch monolingual 10-month-olds (19 girls; M_{age} = 304 days, range 288 – 322 days). Another 20 infants were tested but excluded for the following reasons: inattentive to screen (n = 11); started crying (n = 1); or technical problems, such as being unable to obtain a good calibration (n = 8). Infants were recruited from the babylab database from Utrecht University. All procedures were approved by the local Ethics Committee. All parents signed informed consent before taking part in the experiment. Parents could have their travel costs reimbursed and received a small gift in the form of an age-appropriate book for their child.

3.2.2. Stimuli

The auditory materials were the same as those in Experiment 1. In addition, the experiment included two types of visual stimuli (see Fig. 1). The target visual stimulus during the trials was an eight-by-eight checkerboard (800 × 800 pixels), with some of its white surfaces colored. A total of 63 different checkerboard images were created and alternated pseudorandomly every second. As an attention getter, to be displayed between trials, we used an image of a red spiral, which would rotate slowly around its middle axe in the central position on the screen. Both the checkerboard and the attention getter were surrounded by a gray frame to cover the remainder of the 17" monitor (resolution: 1280 × 1024 pixels).

3.2.3. Design

See Experiment 1.

3.2.4. Apparatus

The experiment was conducted with an arm-mounted EyeLink 1000 eye tracker sampling at 500 Hz with a 940 nm modified illuminator (AM 890) and a 16 mm lens. According to the manufacturer, in remote mode the gaze position accuracy is 0.5 degrees and the range for freedom of movement is 25 by 25 by 10 cm (horizontal x vertical x depth) at a 60 cm distance from the eye tracker. Stimuli were displayed on a 17' Acer AL1717 monitor. Two Tangent EVO E4 speakers (20-100 Watt) inside the experiment room played audio stimuli via a Sony TA-FE230 amplifier at an intensity comparable to Experiment 1 (around 65 dBA). The experiment was programmed in a home-based software program (ZEP) and ran in Linux. It took place in a separate room designed for testing. The experimenter monitored the children via a webcam (Logitech C920 webcam) attached to the top of the screen.

3.2.5. Procedure

Similar to Experiment 1, parents were first informed about the procedure, before they signed the informed consent. The parent

was asked to sit on a chair behind their child, who was seated in a high chair which stood on average 55 cm from the eye tracker. The parent was instructed not to interact with the child once the experiment started and was asked to put on headphones through which calm, instrumental music was played. Subsequently, a calibration procedure with three calibration points took place. The calibration was manually checked by the experimenter and when deemed satisfactory the experiment started.

In the same manner as Experiment 1, the experiment started with the familiarization phase followed by a test phase. All familiarization and test trials displayed images of the checkerboard. Each trial started after a 1-second attention getter (see Fig. 1), accompanied by a sound (musical chimes). This attention getter served a similar function as the flashing of the center light between trials in the HPP experiment. Following the standard procedure of this lab, the 1-second duration of the attention getter was fixed. This stands in contrast to the infant-controlled duration of the center light in Experiment 1, which continued to play until children attended for 1 uninterrupted second. We will return to this potentially important difference between the procedures in the discussion. Trials lasted until completion or until the eye tracker detected fixations outside the area of interest for 2 consecutive seconds or had failed to track the infant's eye for 2 consecutive seconds.

3.2.6. Pre-processing steps

The Area of Interest (AoI) was defined as 900 × 900 pixels, so it was slightly larger than the checkerboard image. All data points (every 2 ms) were aggregated into looks in one of four categories: (1) within the checkerboard (800 × 800 pixels); (2) on the edge of the stimulus, but within the AoI; (3) outside the AoI; (4) blinks or looking away that rendered the eye-tracker unable to track the pupil. Looking time was calculated based on aggregated looks for categories 1 and 2. Looking time data intervals of less than 500 ms were imputed as looks, as long as the child was looking at the screen before as well as after the interval (see also ter Schure, Mandell, Escudero, Raijmakers, & Johnson, 2014). Since the average duration of infant eye blinks is 419 ms (Bacher & Smotherman, 2004), such a criterion allowed us to maintain periods in which infants were blinking. Missing data in intervals longer than 500 ms were coded as a 'look away' from the screen and discarded from total looking time. Trials with less than 1 second of total looking time were removed from the data. Only infants who contributed at least three familiar-word and three novel-word test trials were included in the analyses.

3.3. Experiment 3: Central Fixation procedure using manual coding (CF-M)

3.3.1. Subjects

A final set of 32 full-term, typically developing Dutch monolingual 10-month-olds contributed sufficient data to this third experiment (13 girls; M_{age} = 305 days, range 289 – 318 days). Another 13 infants were tested but excluded for the following reasons: inattentive to screen (n = 2); started crying (n = 8); parental interference (n= 2); or technical problems (n = 1). Infants were recruited from the same babylab database as used for Experiment 2. All procedures were approved by the local Ethics Committee. All parents signed informed consent before taking part in the experiment. Parents could have their travel costs reimbursed and received a small gift in the form of an age-appropriate book for their child.

3.3.2. Stimuli

The auditory speech materials were the same as those in Experiment 1. The visual stimuli and the sound accompanying the attention getter were the same as in Experiment 2.

3.3.3. Design

See Experiment 1.

3.3.4. Apparatus

Testing took place in the same room as in Experiment 2, and we used the same screen and audio equipment presenting the stimuli. The webcam (Logitech C920 webcam) attached to the top of the screen monitored the child's behaviour. The experimenter was seated in an adjacent room, using the real-time webcam feed to manually code infant looking behaviour, while being blind to the audio that was presented to the child. Their button presses served as input to the script that controlled the presentation of trials. The experiment was programmed in a home-based software program (ZEP) and ran in Linux.

3.3.5. Procedure

The procedure was identical to Experiment 2, with two exceptions. First, the experimenter manually coded when the child attended to the screen (see above). Second, as in Experiment 1, but in contrast to Experiment 2, the visual attention getter (see Fig. 1) and accompanying sound were infant-controlled: the attention getter continued to play until the experimenter's button press indicated that the participant had attended for one uninterrupted second. Trials lasted until completion or whenever the experimenter indicated that the child's looking away exceeded 2 seconds.

3.3.6. Pre-processing steps

Similar to Experiment 1, we used ELAN (Sloetjes & Wittenburg, 2008) for off-line coding of children's looking per trial. The same routine was followed as for Experiment 1: Coders were blind to the type of words infants heard and instances of ambiguous behaviour were discussed until researchers reached agreement. In calculating the total looking time from the off-line coding, looks away from the screen were discarded from children's total looking time. Trials with a total looking time of less than 1 second were removed from

the analysis. Infants were only retained for further analyses when they contributed at least three test trials per condition (familiar/novel words).

3.4. Analyses

All analyses were carried out in SPSS 25.0. We first checked whether our three subject populations were comparable in gender distribution (using a Chi-Square test) and in age (using a one-way Analysis of Variance (ANOVA)).

Our primary focus was whether choice of procedure modulated infant preference at test. For this question we combined the three experiments, and analysed experimental outcomes in two ways. First, we conducted a mixed Repeated Measures ANOVA to assess whether the effect of *Word type* (2: familiar words vs. novel words) was modulated by the *Procedure* (3: HPP vs. CF-ET vs. CF-M). Since Shapiro-Wilk tests show significant departures from normality for both the mean looking times (LT) for familiar words ($W(96) = 0.91$, $p < .001$) and those for novel words ($W(96) = 0.93$, $p < .001$), we normalized mean looking times using a Log10 transformation. Since testing procedures differed in the number of test trials (Kruskal-Wallis $H(2) = 26.764$, $p < .001$, cf. Results), we added *Number of valid test trials* as a covariate. In case of a significant interaction effect between *Word type* and *Procedure*, planned post-hoc comparisons contrasted each procedure with the other two procedures, using Bonferroni comparisons. We then also ran *t*-tests for each of the testing procedures to obtain effect sizes.

Second, we entered normalized looking times for test trials into a linear mixed-effects model (LMM) analysis, a form of a General Linear Model that does not assume homogeneity of variance, sphericity or compound symmetry, and, most importantly, allows for missing data points (Goldstein, 2011, 1986; Quené & Van den Bergh, 2008; Snijders, 2011). We ran one analysis with all three experiments in one model, but we also ran separate models for each experiment. In the LMM with all three experiments we considered the variables *Word type* (familiar or novel word), *Procedure* (HPP, CF-ET, or CF-M) as fixed factors, *Subject* as a relevant random factor, *Trial* as a repeated measures factor, and *Number of valid test trials* as a relevant covariate. In the separate models we considered the variables *Word type* (familiar or novel word) as a fixed factor, *Subject* as a relevant random factor, and *Trial* as a repeated measures factor. Significant main effects or interactions are followed up with post-hoc comparisons.

Finally, we compared the three procedures on a range of variables, including number of trials in the familiarization or test phase, and dropout rates. We used one-way ANOVAs or non-parametric tests (Kruskall Wallis test) when one of the variables digressed from a normal distribution. For the dropout comparison we used a Chi-Square test.

4. Results

4.1. Procedure comparisons on subject variables

The experiments do not differ on the male-female ratio per test ($X^2(2, N = 96) = 1.6$, $p = .45$, nor in the subjects' ages ($F(2, 93) < 1$, $p = .78$)

4.2. Procedure comparisons for infant preference at test: Repeated Measures ANOVA

We carried out a Repeated Measures ANOVA on the normalized mean looking times during test with *Word type* as factor, *Procedure* as between-subjects variable, and *Number of test trials* as covariate. While results did not reveal a main effect of *Word type* ($F(1,92) = 3.44$, $p = .07$, $\eta^2_p = .04$) there was a significant interaction between *Word type* and *Procedure* ($F(2,92) = 5.36$, $p = .006$, $\eta^2_p = .10$). The interaction between *Word type* and *Number of test trials* was not significant ($F(1,92) = 3.76$, $p = .06$, $\eta^2_p = .04$).

Planned post-hoc analyses revealed that the HPP differed significantly from the CF-M (HPP vs. CF-M:$t(62) = 3.13$, $p = .006$; HPP vs. CF-ET: $t(62) = 2.17$, $p = .07$; Bonferroni-corrected). The post-hoc analyses did not provide evidence that the eye-tracking and manual versions of the CF procedure differed from each other ($t(62) = 0.80$, $p = .86$).

Paired *t*-tests on the normalised mean looking times for familiar versus novel words revealed a significant familiarity bias for the HPP experiment ($t(31) = 2.72$, $p = .01$): infants listened on average 2.1 seconds ($SD = 4.6$ s; normalized Mean = .10, $SD = .20$) longer to the familiar words than to the novel words. This familiarity bias was displayed by 23/32 infants, which a Wilcoxon signed rank test revealed to be a larger proportion than expected by chance ($Z = -2.41$, $p = .02$).

Neither the CF-ET nor the CF-M yielded significant differences in normalised mean looking times (CF-ET: $M = -0.01$, $SD = 0.19$; $t(31) = -0.30$, $p = .76$; with 15/32 displaying a familiarity bias); and for the CF-M: $M = -0.05$, $SD = 0.16$; $t(31) = -1.61$, $p = .12$; with 13/32 displaying a familiarity bias). See also Fig. 2.

4.3. Procedure comparisons for infant preference at test: Linear Mixed Model (LMM)

For the LMM with all three methods, the model with the best fit was a model with *Subject* as a random factor and *Word type* and *Procedure* and their interaction as fixed factors, and *Number of test trials* as a covariate. *Trial* was considered as a repeated measures factor (within participant), but did not significantly improve the fit of the model. There was no significant main effect of *Word type* ($F(1, 953.41) = .109$, $p = .74$). Neither was there an effect of *Procedure* ($F(2, 93.80) = .456$, $p = .64$), or an interaction effect (*Word type* x *Procedure*) ($F(2, 953.78) = 2.253$, $p = .11$). This suggests that across the three procedures, there was no evidence that infants differentiated familiar words from novel words, nor that their behaviour differed across procedures. *Number of test trials* had a significant effect as a covariate ($F(1, 110.53) = 5.468$, $p = .02$), meaning that children with more valid test trials (>1 s) had on

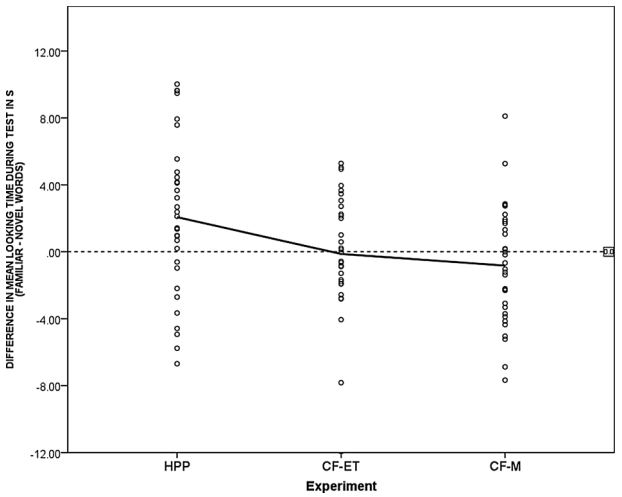

Fig. 2. The mean difference in seconds of looking time at test for familiar minus novel words, split by type of testing procedure. Each circle represents one case. Note: *HPP* = Headturn Preference Procedure; *CF-ET* = Central Fixation procedure with automated Eye Tracking; *CF-M* = Central Fixation procedure with Manual coding. The dotted horizontal line at y = 0 represents no difference in looking times to familiar and novel words; higher scores represent a familiarity preference, and lower scores a novelty preference. The solid horizontal line represents the mean scores for each of the procedures.

average higher mean looking times.

For the LMM for the HPP, the model with the best fit was a model with *Subject* as a random factor and *Word type* as a fixed factor. *Trial* was considered as a repeated measures factor (within participant), but it did not significantly improve the fit of the model. There was no significant main effect of *Word type* ($F(1, 334.48) = 3.770$, $p = .05$).

For the LMM for the CF-ET, the model with the best fit was a model with *Subject* as a random factor and *Word type* as a fixed factor. *Trial* was considered as a repeated measures factor (within participant), but did not significantly improve the fit of the model. There was no significant main effect of *Word type* ($F(1, 288.11) = 0.148$, $p = .70$).

For the LMM for the CF-M, the model with the best fit was a model with *Subject* as a random factor and *Word type* as a fixed factor. *Trial* was considered as a repeated measures factor (within participant), but it did not significantly improve the fit of the model. There was no significant main effect of *Word type* ($F(1, 341.07) = 0\ .997$, $p = .32$).

Table 3 summarizes the estimates of fixed effects outcomes for each of the procedures.

4.4. Procedure comparisons on dropout rates

We also compared the testing procedures on their dropout rates, given that infant studies are generally time-consuming and require a difficult population to recruit. Dropout rates in our experiments are relatively high, which is common in infant research, but it appears that the HPP yields a higher dropout rate (52.9%) than the CF-ET (38.5%) and the CF-M (28.9%). These dropout rates differed significantly from one another ($X^2(2) = 7.535$, $p = .023$). In subsequent method-by-method Chi-Square tests we found that the HPP and the CF-M differed significantly from each other in dropout rate ($X^2(1) = 7.187$, $p = .01$), but the HPP and the CF-ET did not ($X^2(1) = 2.482$, $p = .12$), and neither did the two CF methods ($X^2(1) = 1.343$, $p = .25$). In all three experiments, infants that dropped out did not differ significantly in mean age or sex distribution from included infants. See Table 4.

Dropout in the HPP and CF-ET experiments was mainly due to infants' not attending to, respectively, the lights ($n = 17$; 47%) and the screen in the CF-ET experiment ($n = 10$; 50%). However, only two (16%) infants in the CF-M experiment were inattentive to the screen and the main reason for dropout in this procedure was that infants started crying ($n = 8$; 66%). Crying was the second most frequent reason for dropping out of the HPP ($n = 12$; 33%), but only occurred for two infants (10%) participating in the CF-ET

Table 3
Estimates of fixed effects outcomes for the separate linear mixed-effects models.

	Estimate	Std. Error	*df*	*t*	Sig.	Confidence interval 95%	
						Lower bound	Upper bound
HPP	.0693	.036	334.48	1.942	.053	-.0009	.1394
CF-ET	-.0135	.035	288.11	-.384	.701	-.0828	.0558
CF-M	-.0329	.033	341.07	-.998	.319	-.0977	.0319

Note: HPP = Headturn Preference Procedure; CF-ET = Central Fixation procedure with automated Eye Tracking; CF-M = Central Fixation procedure with Manual coding

Table 4
Overview exclusions vs. inclusions per procedure.

		N	*N* Male	Mean age (SD)
HPP	*Excluded*	36	23	306.25 (8.84)
	Included	32	16	303.09 (10.01)
CF-ET	*Excluded*	20	10	303.75 (8.63)
	Included	32	13	303.53 (9.11)
CF-M	*Excluded*	12	7	305.92 (7.86)
	Included	32	19	304.56 (8.59)

Note: HPP = Headturn Preference Procedure; CF-ET = Central fixation procedure with automated Eye Tracking; CF-M = Central Fixation procedure with Manual coding

experiment. Across experiments, technical problems appeared occasionally, but not notably more in one procedure over another. While for HPP these reflect a variety of reasons, such as human errors or faulty lights, for the CF-ET experiment these errors mainly reflect calibration difficulties, whereas in the CF-M experiment technical errors were also mainly due to human error.

4.5. Procedure comparisons on the familiarization phase

There was no significant difference between experiments in number of trials required to reach familiarization ($F(2,93) = 2.701$, $p = .07$). On average infants in the HPP experiment required listening to 8.8 familiarization trials ($SD = 2.9$; range 5-16), while infants in the CF-ET experiment required 8.3 trials ($SD = 2.0$; range 5-12), and the infants in the CF-M experiment required 7.3 trials ($SD = 2.6$; range 4-14).

4.6. Procedures comparisons on the test phase

The maximum of test trials an infant could listen to was 12. Groups differed in the total number of valid (>1 s.) test trials ($H(2) = 26.764$, $p < .001$). The mean number of valid test trials was 11.4 for the HPP experiment ($SD = 0.8$; range 9-12); 9.95 for the CF-ET experiment ($SD = 1.8$; range 7-12), and 11.7 for the CF-M experiment ($SD = 0.6$; range 10-12). Post-hoc comparisons revealed that the CF-ET procedure yielded fewer trials than the other two procedures (each comparison $p < .001$; Bonferroni-corrected). Note that children with fewer than three valid trials for the target words and three valid trials for the novel words were excluded from analyses of all three experiments, including these analyses on the number of test trials.

Consequently, the testing procedures further differ in the mean total looking time per child during the test phase ($F(2,93) = 4.934$, $p = .009$), with the CF-ET experiment having shorter looking times ($M = 82.17$ s; $SD = 40.7$ s) than either the HPP ($M = 115.44$ s; $SD = 23.3$ s) or the CF-M experiment ($M = 107.82$ s; $SD = 24.1$ s). This difference among experiments likely stems from a smaller number of included test trials for CF-ET.

5. Discussion

This paper compared three different procedures to assess infant preferences in a speech segmentation task: a headturn preference procedure (HPP) and a central-fixation (CF) procedure with either automated eye tracking (CF-ET) or manual coding (CF-M). Across all three procedures, infants were familiarized to words embedded in passages and tested on their looking time to a visual stimulus while hearing lists repeating either these familiar words or novel words. A difference in looking times to the familiar compared to the novel words is regarded as evidence of infants' speech segmentation ability, and longer looking to the familiar compared to the novel words is called a familiarity preference. The key observation of the comparison presented in this study is that the HPP yielded a majority of infants displaying a familiarity preference as well as a group-level familiarity effect, whereas both versions of the CF procedure did not yield a significant looking-time difference between familiar and novel words and thus failed to provide evidence of infants' ability to segment words from passages.

Whether this apparent difference between testing procedures really matters was analysed in two ways, using an Analysis of Variance (ANOVA) on looking times that are averaged over trials as well as a linear mixed-effects model on the by-trial looking times. The field of infant language acquisition has traditionally relied on *t*-tests and (repeated measures) ANOVAs, but the linear mixed-effects model is increasing in popularity as it is considered a more sophisticated and powerful analysis technique that corrects for missing data and correctly accounts for nested data, such as multiple measures from the same child (Goldstein, 1986, 2011). Studies usually conduct and report only one of these statistical analyses and the general assumption seems to be that both should point to the same conclusions when an effect is truly present. In the present case, however, the two analyses yielded different results, which would lead to different conclusions regarding the impact of testing procedure on infants' preferences in a segmentation task. Instead of presenting only one analysis that served our hypothesis (which could be considered *p*-hacking) we deemed it best to report both analyses and illustrate that their results do not always align. The results from the ANOVA suggest that the HPP is more sensitive than both CF procedures in revealing infants' segmentation abilities. However, the results from the linear mixed-effects model reveal no significant differences between the procedures, providing no basis for concluding that the procedures are different. Indeed, there is ample evidence in the literature for each procedure that it is able to expose infant preference. These results are thus inconclusive

regarding the impact that testing procedure has on the researchers' ability to detect infant segmentation ability.

We do note nonetheless that the nominally larger effect size in the HPP is in line with the non-significant trend in our re-analysis of the infant speech segmentation meta-analysis (Bergmann & Cristia, 2016). These results are also in line with the average increased effect size of 0.21 in the HPP compared to the single-screen CF procedure for infants' preference for infant-directed over adult-directed speech (ManyBabies Consortium, 2019). This begs the question why infant preferences might be more pronounced in the HPP than in CF.

One explanation for the apparent difference between the procedures could be that infants were not randomly assigned to the three studies and that the HPP participants were sampled from a population in a different town. However, participants had no influence on the study they participated in and the towns in which the infants were recruited are demographically very similar. We therefore deem this an unsatisfactory explanation.

Another possibility is that some aspect of the HPP renders this procedure intrinsically more sensitive to infants' preferences. Note that the three contrasted procedures were comparable in the contingency between sound and children's attention to an unrelated visual stimulus: the presentation of speech stopped once children looked away for more than 2 seconds. Such contingency does not appear mandatory for revealing infant preferences, as effects can be found in CF studies playing sounds for a fixed duration (e.g., Marquis & Shi, 2008; Shi & Lepage, 2008), but the way the contingency is realized might contribute to the apparently increased sensitivity of the HPP. The present results, together with those of the aforementioned meta-analysis and the ManyBabies Consortium (2019), suggest that infants become more sensitive to the contingency of sounds with their gross motor behaviour, such as headturns, than to the contingency with subtler movements, such as produced by their eyes. Perhaps infants' increased sensitivity to contingency with their own gross motor behaviour results from their daily lives, where infants are likely to have experienced contingent responses to their gross motor acts, such as moving their heads and grasping.

Another speculation is that a child's prior experience with screens interferes with their ability to notice the contingency between sound and their visual behaviour in the CF procedure. Even though most parents indicated that their child had no experience with watching TV, it is likely that most infants would have viewed some videos on a screen without ever experiencing that their looking behaviour influenced the presentation of sounds.

Alternatively, the observed advantage of the HPP might be an artefact that stems from including a less diverse and more optimal subject population. Recall that procedures differed in the number of children excluded as well as the reasons for exclusion. Notably, the HPP stood out as the testing procedure in which most infants were excluded, mainly due to inattentiveness to the lights, followed by crying. Possibly, the HPP is a more challenging procedure for infants, which implies that the infants who successfully complete this procedure may present a cognitively more advanced sample, who are able to display preferences that are still out of reach for their age group as a whole. This argument would be stronger if independent measures of the children's cognitive performance had been taken, which is not the case. However, the present data suggest it is important for all studies to be as clear as possible in accounting for their dropouts and that future comparisons between procedures should continue to explore the impact of dropout rates on result patterns.

The pattern of infant preference that we observed for the HPP was a familiarity preference, while more infants in the CF-M displayed a novelty preference. Regardless of the direction of the preference, any significant difference in looking behaviour for familiar versus novel words has been interpreted as evidence that infants recognized words from the familiarization phase. Although more studies list a familiarity preference, there are cases with a novelty preference, which makes it difficult to predict the direction of preference (Bergmann & Cristia, 2016). Indeed, a recent test-retest study listed very variable results for infants who were tested twice within three days on word segmentation studies (Cristia, Seidl, Singh, & Houston, 2016). This is in marked contrast to predicting the preference for infant-directed versus adult-directed speech, in which case most infants listen longer to the former (Dunst et al., 2012; ManyBabies Consortium, 2019).

Why do we need a better understanding of infant preferences? It becomes increasingly clear that the speech perception skills that infants master in their first year after birth are essential for subsequent language development (Cristia, Seidl, Junge, Soderstrom, & Hagoort, 2014; Kuhl et al., 2008; Werker & Hensch, 2015). This is also the case for speech segmentation ability. In a seminal study, Newman and colleagues showed that infant preference for words presented prior to test over novel words ('familiarity response') was positively related to expressive vocabularies at the age of two (Newman, Ratner, Jusczyk, Jusczyk, & Dow, 2006). Other studies also report positive links between a familiarity response and concurrent as well as subsequent language development, suggesting that those infants who prefer to hear repeated words build larger lexicons (e.g., Singh, Reznick, & Xuehua, 2012; cf. Junge & Cutler, 2014). In contrast, a recent study argues that the infants showing a preference for novel words ('novelty response') will be more linguistically advanced (Depaolis, Vihman, & Keren-Portnoy, 2014; Newman, Rowe, & Ratner, 2016;). These discrepancies reveal that difficulties predicting infant preference patterns also lead to inconsistencies in determining which patterns of infant responses are predictive of advanced language development.

Of course, there are alternatives to behavioral procedures to ascertain whether infants can discriminate between two types of stimuli. For instance, electrophysiological measurements of infant word segmentation indexing only recognition of familiar words (and not necessarily preference) also correlate with lexical development at two years of age (Junge, Kooijman, Hagoort, & Cutler, 2012). However, electrophysiological recordings are not always available to infant researchers. Moreover, while neuroimaging techniques are becoming more popular, the majority of infant studies rely on infant preferences to assess discrimination (Oakes, 2012).

A final contribution of this paper is the comparison between the procedures on a range of other variables. Infants in the CF-ET contributed on average fewer test trials than infants in the other two experiments. Since this difference was also apparent in comparison with the CF-M, which was highly similar in its presentation to infants, we can rule out the possibility that the checkerboard

stimulus is not sufficiently appealing. One likely factor that explains the increased trial attrition rates in the CF-ET is that the attention getter between trials was of a fixed duration, such that trials might have started without the infant fixating the visual display. Hence, we advise researchers to build in the option for test trials to only begin when the child is attentive. Another potential explanation for the smaller number of trials in the CF-ET is related to the use of an automatic eye tracker. Such means of tracking infants' eyes usually become less precise as a function of time, revealing increasingly larger periods of data loss (Hessels et al., 2015). This data loss will be interpreted as a look away from the screen, but could instead reflect a failure in the eye tracker to track the child's eyes because she has moved her head away from the position in which she was originally calibrated. One solution to overcome such data loss would be to recalibrate whenever the experimenter sees that the child is attending to the screen, while the eye tracker registers data loss.

To conclude, the present study has demonstrated that there is some reason to believe that the headturn preference procedure is more effective than the central-fixation procedure at detecting infant preferences in learning paradigms, such as the ones used in speech segmentation studies.

However, the evidence from this study is not very strong and it seems premature to rely on this outcome to guide the interpretation of past work or the decision-making regarding the configuration of future babylabs. The dropout rates across the three studies suggest that a high dropout rate may be one factor that contributes to the increased sensitivity of the headturn preference procedure, which raises questions about the generalizability of the results. We believe it is through cross-laboratory collaborations such as the ManyBabies Consortium (2019; ManyBabies Consortium (2019; see also Cristia et al., 2016) and storing data in online repositories for future inspection (e.g., as in Metalab, 2020: Bergmann et al., 2018) that we can start revealing the many variables that contribute to exposing infant preference for speech.

6. Term Definition

Caroline Junge (CJ), Emma Everaert (EE), Lyan Porto (LP), Paula Fikkert (PF), Maartje de Klerk (MdK), Brigitta Keij (BK), and Titia Benders (TB)

Conceptualization CJ, EE, and TB
Methodology All Authors
Software TB, EE, LP
Validation TB
Formal analysis EE, CJ, LP, TB
Investigation CJ, EE, LP, BK
Resources PF, MdK
Data Curation LP, TB, EE, CJ
Writing - Original Draft CJ, EE
Writing - Review & Editing CJ, EE, LP, PF, TB
Visualization CJ
Supervision CJ PF BK
Project administration CJ
Funding acquisition CJ

Division of labour: CJ, TB, EE, BK, MdK and PF designed the experiments. CJ, LP and TB created the stimuli. LP, TB, CJ and PF ran the first experiment, while EE, BK and CJ conducted the second experiment, and CJ, EE, MdK, and TB were responsible for the third study. CJ, EE, LP, and TB analyzed the data. CJ, EE, LP, PF and TB wrote the paper.

Acknowledgements

The authors would like to thank parents and infants for their cooperation. We thank Imme Lammertink and Maaike van Buren for assistance in the Headturn Preference Procedure Experiment; Charlotte Koevoets and Danique van Aalst for their assistance in the manual version of the Central Fixation Procedure, and Susanne Brouwer for lending her voice. We gratefully acknowledge Christina Bergmann for helping us with the meta-analyses. We also express thanks to the editor and anonymous reviewers for their constructive feedback. The first author was supported by a personal VENI grant (016.154.051) from the Dutch Organization for Scientific Research (NWO).

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.infbeh.2020.101448.

References

Aslin, R. N. (2007). What's in a look? *Developmental science, 10*(1), 48–53.
Aslin, R. N. (2012). Infant eyes: A window on cognitive development. *Infancy, 17*(1), 126–140. https://doi.org/10.1111/j.1532-7078.2011.00097.x.
Bacher, L. F., & Smotherman, W. P. (2004). Spontaneous eye blinking in human infants: A review. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology, 44*(2), 95–102.
Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: a meta-analytic approach. *Developmental science, 19*(6), 901–917.

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., ... Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89*(6), 1996–2009. https://doi.org/10.1111/cdev.13079.
Best, C., & Jones, C. (1998). Stimulus-alternation preference procedure to test infant speech discrimination. *Infant Behavior and Development,* (21), 295.
Boersma, P., & Weenink, D. (2015). *Praat: doing phonetics by computer [Computer program]* (Version 5.4.18). Retrieved September 7, 2015, from http://www.praat.org/.
Cooper, R. P., & Aslin, R. N. (1990). Preference for infant-directed speech in the first month after birth. *Child development, 61*(5), 1584–1595. https://doi.org/10.1111/j.1467-8624.1990.tb02885.x.
Cristia, A., Seidl, A., Junge, C., Soderstrom, M., & Hagoort, P. (2014). Predicting individual variation in language from infant speech perception measures. *Child development, 85*(4), 1330–1345.
Cristia, A., Seidl, A., Singh, L., & Houston, D. (2016). Test–retest reliability in infant speech perception tasks. *Infancy, 21*(5), 648–667.
Depaolis, R. A., Vihman, M. M., & Keren-Portnoy, T. (2014). When do infants begin recognizing familiar words in sentences? *Journal of child language, 41*(1), 226–239. https://doi.org/10.1017/S0305000912000566.
Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning, 5*(1), 1–13.
Fernald, A. (1985). Four-month-old infants prefer to listen to motherese. *Infant behavior and development, 8*(2), 181–195. https://doi.org/10.1016/S0163-6383(85)80005-9.
Floccia, C., Keren-Portnoy, T., DePaolis, R., Duffy, H., Delle Luche, C., Durrant, S., ... Vihman, M. (2016). British English infants segment words only with exaggerated infant-directed speech stimuli. *Cognition, 148*, 1–9. https://doi.org/10.1016/j.cognition.2015.12.004.
Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... Lew-Williams, C. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy, 22*(4), 421–435.
Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*(1), 43–56.
Goldstein, H. (2011). *Multilevel statistical models*, *Vol. 922*. Hoboken, NJ: John Wiley & Sons.
Gredebäck, G., Johnson, S., & von Hofsten, C. (2009). Eye tracking in infancy research. *Developmental neuropsychology, 35*(1), 1–19. https://doi.org/10.1080/87565640903325758.
Hessels, R. S., Andersson, R., Hooge, I. T., Nyström, M., & Kemner, C. (2015). Consequences of Eye Color, Positioning, and Head Movement for Eye-Tracking Data Quality in Infant Research. *Infancy, 20*(6), 601–633. https://doi.org/10.1111/infa.12093.
Hessels, R. S., & Hooge, I. T. (2019). Eye tracking in developmental cognitive neuroscience–The good, the bad and the ugly. *Developmental cognitive neuroscience, 40*, 100710.
Houston, D. M., Jusczyk, P. W., Kuijpers, C., Coolen, R., & Cutler, A. (2000). Cross-language word segmentation by 9-month-olds. *Psychonomic Bulletin & Review, 7*(3), 504–509. https://doi.org/10.3758/BF03214363.
Junge, C., & Cutler, A. (2014). Early word recognition and later language skills. *Brain sciences, 4*(4), 532–559.
Junge, C., Kooijman, V., Hagoort, P., & Cutler, A. (2012). Rapid recognition at 10 months as a predictor of language development. *Developmental Science, 15*(4), 463–473.
Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive psychology, 29*(1), 1–23.
Jusczyk, P. W., Houston, D. M., & Newsome, M. (1999). The beginnings of word segmentation in English-learning infants. *Cognitive psychology, 39*(3), 159–207.
Kooijman, V., Hagoort, P., & Cutler, A. (2009). Prosodic structure in early word segmentation: ERP evidence from Dutch ten-month-olds. *Infancy, 14*(6), 591–612.
Kuhl, P. K., Conboy, B. T., Coffey-Corina, S., Padden, D., Rivera-Gaxiola, M., & Nelson, T. (2008). Phonetic learning as a pathway to language: new data and native language magnet theory expanded (NLM-e). *Philosophical Transactions of the Royal Society B: Biological Sciences, 363*(1493), 979–1000.
ManyBabies Consortium (2019). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science*.
Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition, 82*(3), B101–B111.
Marquis, A., & Shi, R. (2008). Segmentation of verb forms in preverbal infants. *The Journal of the Acoustical Society of America, 123*(4), EL105–EL110. https://doi.org/10.1121/1.2884082.
Meints, K., & Woodford, A. (2008). *Lincoln Infant Lab Package 1.0: A new programme package for IPL, Preferential Listening, Habituation and Eyetracking.[www document: Computer software & manual]*. http://www.lincoln.ac.uk/psychology/babylab.htm.
Metalab Interactive tools for community-augmented meta-analysis, power analysis and experimental planning in cognitive development research. Data downloaded on 2019, December 19. Retrieved from http://metalab.stanford.edu.
Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant behavior and development, 16*(4), 495–500.
Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language, 43*(1), 1–19.
Nazzi, T., Mersad, K., Sundara, M., Iakimova, G., & Polka, L. (2014). Early word segmentation in infants acquiring Parisian French: task-dependent and dialect-specific aspects. *Journal of Child Language, 41*(3), 600–633. https://doi.org/10.1017/S0305000913000111.
Newman, R. S., Ratner, N. B., Jusczyk, A. M., Jusczyk, P. W., & Dow, K. A. (2006). Infants' early ability to segment the conversational speech signal predicts later language development: a retrospective analysis. *Developmental Psychology, 42*(4), 643–655.
Newman, R. S., Rowe, M. L., & Ratner, N. B. (2016). Input and uptake at 7 months predicts toddler vocabulary: the role of child-directed speech and infant processing skills in language development. *Journal of child language, 43*(5), 1158–1173.
Oakes, L. M. (2012). Advances in eye tracking in infancy research. *Infancy, 17*(1), 1–8. https://doi.org/10.1111/j.1532-7078.2011.00101.x.
Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*(4), 413–425. https://doi.org/10.1016/j.jml.2008.02.002.
Schreiner, M. S., & Mani, N. (2017). Listen up! Developmental differences in the impact of IDS on speech segmentation. *Cognition, 160*, 98–102.
Shi, R., & Lepage, M. (2008). The effect of functional morphemes on word segmentation in preverbal infants. *Developmental Science, 11*(3), 407–413. https://doi.org/10.1111/j.1467-7687.2008.00685.x.
Shi, R., Werker, J. F., & Cutler, A. (2006). Recognition and representation of function words in English-learning infants. *Infancy, 10*(2), 187–198.
Singh, L., Reznick, S. J., & Xuehua, L. (2012). Infant word segmentation and childhood vocabulary development: a longitudinal analysis. *Developmental science, 15*(4), 482–495.
Sloetjes, H., & Wittenburg, P. (2008). Annotation by category-ELAN and ISO DCR. *6th international Conference on Language Resources and Evaluation (LREC 2008)*.
Snijders, T. (2011). Multilevel Analysis. In M. Lovric (Ed.). *International Encyclopedia of Statistical Science* (pp. 879–882). New York, NY: Springer.
Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature, 388*(6640), 381.
Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Developmental science, 8*(5), 432–443.
ter Schure, S. M. M., Junge, C. M. M., & Boersma, P. P. G. (2016). Semantics guide infants' vowel learning: computational and experimental evidence. *Infant Behavior and Development, 43*, 44–57.
ter Schure, S. M. M., Mandell, D. J., Escudero, P., Raijmakers, M. E., & Johnson, S. P. (2014). Learning Stimulus-Location Associations in 8-and 11-Month-Old Infants: Multimodal Versus Unimodal Information. *Infancy, 19*(5), 476–495.
Tsuji, S., Fikkert, P., Yamane, N., & Mazuka, R. (2016). Language-general biases and language-specific experience contribute to phonological detail in toddlers' word representations. *Developmental psychology, 52*(3), 379–390. https://doi.org/10.1037/dev0000093.
van Heugten, M., & Johnson, E. K. (2012). Infants exposed to fluent natural speech succeed at cross-gender word recognition. *Journal of Speech, Language, and Hearing Research, 55*(2), 554–560. https://doi.org/10.1044/1092-4388.
Werker, J. F., & Hensch, T. K. (2015). Critical periods in speech perception: new directions. *Annual review of psychology,* 173–196. https://doi.org/10.1146/annurev-psych-010814-015104.
Yoshida, K. A., Pons, F., Maye, J., & Werker, J. F. (2010). Distributional phonetic learning at 10 months of age. *Infancy, 15*(4), 420–433.