Cognition 213 (2021) 104757

Contents lists available at ScienceDirect

Cognition

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

Quantifying the role of rhythm in infants' language discrimination abilities: A meta-analysis☆

Loretta Gasparini [a,*,1], Alan Langus [b], Sho Tsuji [c], Natalie Boll-Avetisyan [b]

[a] European Master's in Clinical Linguistics+ (EMCL+, University of Groningen, University of Potsdam, University of Eastern Finland), Harmoniebuilding 1315, P. O. Box 716, Groningen 9700AS, the Netherlands
[b] Department of Cognitive Sciences, University of Potsdam, House 14, Karl-Liebknecht-Straße 24-25, 14476 Potsdam, Germany
[c] International Research Center for Neurointelligence, The University of Tokyo, 7-3-1 Hongo Bunkyo-ku, Tokyo 113-0033, Japan

ARTICLE INFO

Keywords:
Language discrimination
Accent discrimination
Speech rhythm
Durational cues
Infant speech perception
Meta-analysis

ABSTRACT

More than 30 years have passed since Mehler et al. (1988) proposed that newborns can discriminate between languages that belong to different rhythm classes: stress-, syllable- or mora-timed. Thereupon they developed the hypothesis that infants are sensitive to differences in vowel and consonant interval durations as acoustic correlates of rhythm classes. It remains unknown exactly which durational computations infants use when perceiving speech for the purposes of distinguishing languages. Here, a meta-analysis of studies on infants' language discrimination skills over the first year of life was conducted, aiming to quantify how language discrimination skills change with age and are modulated by rhythm classes or durational metrics. A systematic literature search identified 42 studies that tested infants' (birth to 12 months) discrimination or preference of two language varieties, by presenting infants with auditory or audio-visual continuous speech. Quantitative data synthesis was conducted using multivariate random effects meta-analytic models with the factors rhythm class difference, age, stimulus manipulation, method, and metrics operationalising proportions of and variability in vowel and consonant interval durations, to explore which factors best account for language discrimination or preference. Results revealed that smaller differences in vowel interval variability (△V) and larger differences in successive consonantal interval variability (rPVI-C) were associated with more successful language discrimination, and better accounted for discrimination results than the factor rhythm class. There were no effects of age for discrimination but results on preference studies were affected by age: the older infants get, the more they prefer non-native languages that are rhythmically similar to their native language, but not non-native languages that are rhythmically distinct. These findings can inform theories on language discrimination that have previously focussed on rhythm class, by providing a novel way to operationalise rhythm in language in the extent to which it accounts for infants' language discrimination abilities.

1. Introduction

In 1988, Jacques Mehler and his colleagues found that newborns and two-month-old infants were able to discriminate between certain languages and not others (Mehler et al., 1988). This work was soon extended by the finding that despite their limited experience with

Abbreviations: AIC, Akaike Information Criterion; CF, Central Fixation; CAMA, Community Augmented Meta-analysis; CI, Confidence Interval; EEG, Electroencephalography; FC, Forced Choice; HAS, High-Amplitude Sucking; HPP, Head-turn Preference Procedure; LPF, Low-Pass Filtered; LRT, Likelihood Ratio Test; NIRS, Near-infrared spectroscopy; nPVI-V, Normalised Pairwise Variability Index for vowels; rPVI-C, Raw Pairwise Variability Index for consonants; PRISMA, Preferred Reporting Items for Systematic review and Meta-Analysis; SD, Standard Deviation; SE, Standard Error; UC, Unclassified; VarcoV, Standard deviation of vocalic interval durations divided by the mean; VIF, Variance Inflation Factors; △C, Standard deviation of consonantal interval durations; △V, Standard deviation of vocalic interval durations; %V, Vocalic interval durations as percentage of utterance duration.

☆ This paper is a part of special issue "Special Issue in Honour of Jacques Mehler, Cognition's founding editor".

* Corresponding author.
E-mail addresses: gasparini.lorett@gmail.com (L. Gasparini), alan.langus@uni-potsdam.de (A. Langus), shotsuji@ircn.jp (S. Tsuji), nboll@uni-potsdam.de (N. Boll-Avetisyan).
[1] Present postal address: Murdoch Children's Research Institute, Royal Children's Hospital, Flemington Road, Parkville Victoria 3052 Australia

https://doi.org/10.1016/j.cognition.2021.104757
Received 31 August 2020; Received in revised form 13 April 2021; Accepted 27 April 2021
Available online 24 May 2021

speech, newborns already prefer their native language over a non-native one (Moon, Cooper, & Fifer, 1993). Mehler and colleagues' (1988) initial proposal was that newborns' successful discrimination depended on their familiarity with one of the tested languages. Considering that foetuses perceive suprasegmental, prosodic (rhythmic) information contained in the speech signal in the womb, Mehler, Dupoux, Nazzi, and Dehaene-Lambertz (1996) hypothesised that babies begin life with a sensitivity to the rhythmic properties of their native language, and thus do not discriminate between all languages, but between those that belong to different rhythm classes. The three traditionally proposed rhythm classes are stress-, syllable- and mora-timed, classified based on whether the basic prosodic unit that is believed to repeat at regular intervals is the inter-stress interval, syllable or mora (Abercrombie, 1967). Later, Mehler and colleagues used an acoustically-defined approach of operationalising linguistic rhythm, proposing that infants are sensitive to differences in consonant and vowel interval durations between languages if these are sufficiently distinct, and suggesting that patterns in these durational cues create the sensation of rhythm class (Ramus, Nespor, & Mehler, 1999). However, many infant language discrimination studies have continued to consider the role of rhythm class without turning to a more acoustically-based definition of rhythmic differences between languages (Nácar Garcia, Guerrero-Mosquera, Colomer, & Sebastián-Gallés, 2018; White, Luche, & Floccia, 2016), and only a few, more recent studies have directly tested the role of durational metrics on infants' language preferences (Paillereau, Podlipský, Smolík, Šimáčková, & Chládková, 2021) or discrimination abilities (White, Floccia, Goslin, & Butler, 2014, see Section 1.1). The goal of the present meta-analysis is to synthesise the current evidence on infants' language discrimination abilities (including dialects, closely-related varieties, and accents, lexically and morpho-syntactically equivalent varieties with phonetic or prosodic differences), thereby probing Mehler and colleagues' hypotheses on the development of infants' sensitivity to rhythm-based language discrimination, and the relative importance of rhythm class, language familiarity, and bottom-up durational cues therein.

Initial evidence supporting Mehler and colleagues' (1996) proposals regarding the role of rhythm in language discrimination came from follow-up studies showing that young infants could discriminate between two non-familiar languages, as long as they were rhythmically distinct, defined as belonging to different rhythm classes (Christophe & Morton, 1998; Nazzi, Bertoncini, & Mehler, 1998). These results indicate that surface rhythmic cues are available to infants, at least at very young ages, for language discrimination, regardless of their familiarity with the languages under question. The fact that infants rely on rhythmic information rather than segmental information when discriminating between languages has been supported by results from experiments that have used low-pass filtered or resynthesised stimuli and shown that prosodic (duration, pitch and amplitude) cues tend to be sufficient for infants to discriminate languages (Bosch & Sebastián-Gallés, 1997; Byers-Heinlein, Burns, & Werker, 2010; Molnar, Gervain, & Carreiras, 2013). These studies on infants' ability to rely on rhythmic information for distinguishing between languages later became crucial building blocks in the development of prominent language acquisition theories, such as the prosodic bootstrapping theory (Gleitman & Wanner, 1982; Morgan & Demuth, 1996; Weissenborn & Höhle, 2001), which posits sensitivity to rhythm as a precursor to infants' continuous speech segmentation skills and acquisition of morphosyntax, also known as rhythmic segmentation (Abboub, Boll-Avetisyan, Bhatara, Höhle, & Nazzi, 2016; Butler & Frota, 2018; Nazzi, Iakimova, Bertoncini, Fredonie, & Alcantara, 2006; White, Benavides-Varela, & Mády, 2020; for reviews see Gervain, 2018; Gervain & Mehler, 2010; Langus, Mehler, & Nespor, 2017; Nazzi & Ramus, 2003).

Numerous follow-up studies investigated the interplay of language familiarity and rhythm in the development of infants' language discrimination abilities and preference as they get older and acquire new information about their native language. Babies seem to become familiar enough with their native language that they can discriminate it from any non-native one by sometime between five months (Bosch & Sebastián-Gallés, 1997; Nazzi, Jusczyk, & Johnson, 2000; Zacharaki & Sebastian-Galles, 2021) and seven months (Chong, Vicenik, & Sundara, 2018).

Two alternative theories, either of which could account for heretofore mentioned findings, are the *rhythmic class acquisition hypothesis* (Nazzi et al., 1998), and the *native language acquisition hypothesis* (Nazzi et al., 2000). The *rhythmic class acquisition hypothesis* posits that with time, infants become experts in the common rhythmic organisation of their native rhythmic class. Thus, it predicts that babies are sensitive to the rhythmic properties of languages in their native rhythm class, native and non-native alike, and so should be able to discriminate between two non-native languages in their native rhythm class. Supporting these predictions, Johnson and Braun (2011) found that English-learning 4.5-month-olds could discriminate between non-native German and Norwegian (although they attributed their findings to intonation, not rhythmic, differences). The *native language acquisition hypothesis* emphasises infants' acquisition of rhythmic and other information about their native language, not their native rhythm class, which helps them to discriminate between languages. The latter hypothesis posits that infants are sensitive to (rhythmic) properties of their native language, but not to those of non-native languages, regardless of whether they are in the native rhythm class, and so predicts that babies should not discriminate between two non-native languages in the native rhythm class. Supporting this latter hypothesis, Nazzi et al. (2000) found English-learning 5-month-olds could not discriminate between non-native German and Dutch. Based on extant studies, whether infants primarily attune to their native language or native rhythm class early in life remains an open question.

Other studies have investigated whether these hypotheses also extend to infants' ability to discriminate between different dialects or accents of the same language. One accent discrimination study found that 5-month-olds discriminate non-native accents from their native variety, but cannot discriminate between two non-native accents even if they are acoustically further apart (Butler, Floccia, Goslin, & Panneton, 2011). This supports the *native language acquisition hypothesis*, suggesting that after a certain period of exposure, infants become sensitive to differences relative to their native variety, as opposed to discerning objective differences in acoustic cues. Infants may also lose the ability to perceive within-native-language accent differences sometime between six and nine months of age (Kitamura, Panneton, & Best, 2013).

Another area of research has assessed whether these hypotheses would extend to bilingual infants. Multilingual infants need to discriminate between two or more languages that they regularly hear in their environment (see Höhle et al., 2020, for a review), and have been shown to already possess this ability as newborns, at least if the languages are rhythmically distinct (Byers-Heinlein et al., 2010). Discrimination between two native rhythmically similar languages has been found in bilingual infants aged three and four months (Bosch & Sebastián-Gallés, 2001; Molnar et al., 2013). In general, the presence or absence of the ability to discriminate between languages based on rhythm and familiarity does not seem to differ between monolinguals and bilinguals, but bilinguals may show increased attention to their native languages at least between 3 and 5 months (Molnar et al., 2013; Nácar Garcia et al., 2018). In the present meta-analysis, we include studies on monolingual and bilingual populations, to obtain a comprehensive understanding of the role that rhythm plays in influencing infants' language discrimination skills and preferences.

1.1. Rhythm classes and acoustic correlates of rhythm

Mehler and colleagues' *rhythmic class acquisition hypothesis* made reference to the traditional theories of speech rhythm (Abercrombie, 1967; Bloch, 1950; Ladefoged, 1975; Pike, 1945) that have since turned out not to be empirically supported (Borzone de Manrique & Signorini,

1983; Dauer, 1983; Roach, 1982; Wenk & Wioland, 1982; for more recent reviews challenging rhythm classes see also Nolan & Jeon, 2014; Turk & Shattuck-Hufnagel, 2013; White & Malisz, 2020). Hence, Ramus et al. (1999) hypothesised that there must be acoustic surface correlates of speech rhythm that infants use for distinguishing between languages. They proposed conceiving of linguistic rhythm as proportions of consonant versus vowel interval durations (%V, vocalic interval durations as percentage of utterance duration) and variability in consonant and vowel interval durations (ΔC and ΔV, the standard deviation of consonantal and vowel interval durations respectively). They proposed from their analysis of eight languages that a combination of %V and △C was the clearest acoustic correlate of the traditional rhythm classes.

This work was followed up by Grabe and Low (2002) who introduced Pairwise Variability Indices as novel metrics of speech rhythm. These metrics operationalise interval durational variability as the mean of durational differences between successive intervals (rPVI-C, the mean of differences between successive consonant intervals; nPVI-V, the mean of differences between successive intervals divided by their sum (x100)). They argued that these measures are more robust in the face of between-speaker and -utterance variability than △C and △V. White and Mattys (2007) proposed an alternative metric of vowel interval variability normalised for speech rate (VarcoV, standard deviation of vocalic interval durations, divided by the mean (×100)), proposing that the combination of VarcoV and %V was the best acoustic correlate of traditional rhythm class.

Despite the move towards an account of linguistic rhythm based on the duration and variability of segmental units, the aforementioned studies that proposed durational metrics do not offer a departure from the concept of rhythm class, since they assess these metrics with the goal of investigating how these help to cluster languages into rhythm classes, even when they find that not all languages straightforwardly fall into one of the three traditional rhythm classes (Grabe & Low, 2002; see also Boll-Avetisyan et al., 2020). Perhaps partly for this reason, the concept of rhythm class has persisted in recent developmental literature, as means of explaining which languages infants can discriminate between (e.g., Höhle et al., 2020; Nácar Garcia et al., 2018). Meanwhile, in the phonetics literature, rhythm class is generally accepted as having no empirical basis (Turk & Shattuck-Hufnagel, 2013).

If infants can discriminate between languages (largely) based on the variability in segment durations, as the *rhythmic class acquisition hypothesis* suggests, it is pertinent to find out exactly which durational computations infants use when perceiving speech, and how this can be most accurately operationalised as durational metrics. In the present study we investigate the extent to which a selection of six most widely applied durational metrics (see Table 1; Grabe & Low, 2002; Ramus et al., 1999; White & Mattys, 2007), as compared to rhythm class, account for infants' language discrimination abilities.

1.2. Acoustic cues predicting language discrimination

There is little consensus on which acoustic cues account for infants' ability to discriminate between some languages and not others. Regarding the importance of vowel versus consonant interval variability, early predictions were that infants would particularly focus on vowel interval durations (Mehler et al., 1996; Nespor, Peña, & Mehler, 2003; Ramus et al., 1999), since vowels carry more energy in the speech signal and newborns have been found to pay more attention to vowels than consonants for detecting phonetic differences between syllables (Bertoncini, Bijeljac-Babic, Jusczyk, Kennedy, & Mehler, 1988). Recent converging evidence has shown that vowels can be perceived prenatally more readily than consonants (Moon, Lagercrantz, & Kuhl, 2013), and that newborns pay more attention to vowels for remembering words (Benavides-Varela, Hochmann, Macagno, Nespor, & Mehler, 2012; Nazzi & Cutler, 2019). However, the question of whether it is specifically vowel or consonant interval durations (or both) that predict infants' language discrimination has only rarely been directly investigated in infant language discrimination studies.

The hitherto sole studies that directly tested this question by probing which durational metrics (measures of vowel and consonant interval duration, proportions and variability; speech rate; utterance-final lengthening) best predict discrimination in adults (White, Mattys, & Wiget, 2012) and infants (White et al., 2014) found the following: adults' discrimination of resynthesised Spanish and English normalised for speech rate was best predicted by the raw Pairwise Variability Index for consonants (rPVI-C), suggesting they were sensitive to variability in successive consonantal interval durations, as well as a measure of utterance-final vocalic lengthening (nFinal-V, which we do not analyse in the current study due to insufficient data; White et al., 2012). Speech rate became the best predictive cue when stimuli were not normalised for speech rate. Meanwhile, when discriminating between two varieties that are similar in their interval durations and stress distribution (Welsh Valleys and Orkney English), utterance-final lengthening arose as a timing cue recruited for language discrimination. Utterance-final lengthening was also found to predict infants' discrimination of various accents of English (White et al., 2014). To our knowledge, this has not been investigated in infants in two language varieties that are more distinct, which could reveal whether rPVI-C arises as a predictor of discrimination as it did in adults.

Table 1
Durational metrics relevant for language discrimination.

%V	vocalic interval durations as percentage of utterance duration
△C	standard deviation of consonantal interval durations
△V	standard deviation of vocalic interval durations
VarcoV	standard deviation of vocalic interval durations, divided by the mean (×100) (normalised for speech rate)
nPVI-V	normalised Pairwise Variability Index for vowels. Mean of differences between successive intervals divided by their sum (×100) (normalised for speech rate)
rPVI-C	raw Pairwise Variability Index for consonants. Mean of differences between successive intervals

1.3. Rationale for the current meta-analysis

According to the *rhythmic class acquisition hypothesis* and the *native language acquisition hypothesis*, rhythm is a vital cue for language discrimination early in life, the importance of which attenuates as infants tune in to their native language (Nazzi & Ramus, 2003). With the current body of evidence, the exact developmental trajectory of the role of rhythm in language discrimination is still unclear. Since the publication of the seminal paper by Mehler et al. (1988), many researchers have investigated language discrimination in various age ranges, languages and rhythm classes, using different experimental methodologies (e.g. high-amplitude sucking, head-turn preference procedure, central fixation and neurophysiological measures) and speech manipulations. Other studies have investigated language preferences in infants (Dehaene-Lambertz & Houston, 1998; Moon et al., 1993), which is informative because showing a preference entails the ability to discriminate (but showing no preference does not preclude discrimination, e.g., Byers-Heinlein et al., 2010). To better understand the exact role of rhythm cues in infants' early speech perception, the present study aims at offering a synthesis of the existing body of evidence by conducting a systematic review and meta-analysis. Meta-analysis involves calculating standardised effect sizes to quantitatively synthesise the available body of evidence, which yields greater power and precision than appraising individual studies separately (Cristia et al., 2020). By establishing the effects of age and other relevant moderators, the developmental trajectory of infants' language discrimination skills and the role of rhythm over the first year of life can be presented based on the existing body of evidence as a whole.

Here, we take two approaches in operationalising rhythm in languages. First, we evaluate infants' language discrimination abilities as a

function of whether the tested languages belong to the same of the three traditional rhythm classes, or to different rhythm classes. The second approach is to evaluate language discrimination as a function of the tested languages' distance in durational metrics that characterise systematic surface timing patterns at the segmental level (Grabe & Low, 2002; Ramus et al., 1999; White et al., 2012; White et al., 2014). We then compare these two approaches, to establish which better accounts for infants' language discrimination abilities over the first year of life.

1.4. Objectives

The objectives of the systematic review and meta-analysis are as follows. The first objective is to systematically summarize the current evidence on infant language (including dialect and accent) discrimination abilities. Secondly, we aim to estimate effect sizes and the impact of the rhythmic properties of tested languages (defined either as rhythm class, or by durational metrics) at different ages, as well as the role of methodological factors (i.e., to assess potential effects of experimental paradigm). Third and finally, we strive to create a Community-Augmented Meta-Analysis (Cristia et al., 2020; Tsuji, Bergmann, & Cristia, 2014) of language discrimination studies, which involves the full dataset being publicly available to view on MetaLab (Bergmann et al., 2018; see Section 4.4), which will continuously be updated with new results from future studies. In conducting this systematic review, we have two research questions, the first targeting the classical assumption (following Christophe & Morton, 1998; Nazzi et al., 1998) that infants are sensitive to rhythm classes; the second targeting the assumption (following Ramus et al., 1999) that infants attend to acoustic surface cues to rhythm.

Research Question 1: How do typically-developing infants' abilities to discriminate between languages in the same or different rhythm classes change from birth up to 12 months of age? *Hypothesis:* Based on previous research (Christophe & Morton, 1998; Nazzi et al., 1998), we test the hypothesis that young infants can discriminate between any language varieties traditionally defined as rhythmically distinct (i.e., that fall into different "rhythm classes"). According to the *native language acquisition hypothesis*, as infants get older, they should increasingly discriminate in relation to their native language, and native accent, while the *rhythmic class acquisition hypothesis* posits that with age, infants increasingly become experts in their native rhythm class. *Prediction:* Effect sizes are expected to be larger, especially in younger infants, for any between-rhythm-class contrasts compared to within-rhythm-class contrasts.

Research Question 2: Which durational cue(s) best predict infants' language discrimination skills from newborns up to 12 months of age? *Hypothesis:* Following Ramus et al. (1999), we test the hypothesis that young infants are sensitive to acoustic surface cues that point to differences in "rhythm class", which enables them to discriminate between languages that are rhythmically distinct. Specifically, we suggest that very young infants are sensitive to absolute differences in consonant and vowel interval durations between languages, then as they get older, they become increasingly sensitive to various acoustic differences relative to their native language(s) and thus with age rely less on rhythmic information. *Prediction:* Effect sizes of language discrimination, especially in younger infants, are expected to be predicted by differences in consonantal and vocalic interval durational metrics. In this analysis we explore which combination of these are the best predictors and how effect sizes change with age.

2. Method

We follow the guidelines of the Preferred Reporting Items for Systematic review and Meta-Analysis (PRISMA, Liberati et al., 2009; Moher et al., 2009, see Appendix A). A protocol including planned methods and analyses was published on Open Science Framework (https://osf.io/396yb/) on April 16th, 2020, with later amendments to the proposed method indicated and timestamped. All identified documents, screening and inclusion decisions, the full dataset of included studies, and reproducible code of the quantitative analyses are also available on the OSF page. Future updates and additions to the dataset will be reported in the MetaLab CAMA (http://metalab.stanford.edu/dataset/langdiscrim/).

2.1. Eligibility criteria

Relevant studies were identified based on the following inclusion criteria:

(i) Discrimination or preference between two languages, dialects or accents was the key component of the task.
(ii) The dependent variable was a difference in response to stimuli in two different language varieties.
(iii) Participants were infants aged from 0 days to 11 months, 31 days.
(iv) Participants were typically-developing, born at full-term, with no visual or hearing impairments.
(v) Stimuli were derived from continuous, natural speech, presented in the auditory modality. Single sounds, syllables or words, words lists, or backward speech were not eligible. Audio-visual stimuli were allowed if the auditory component fulfilled the above criteria, and the video was consistently included and congruent with the audio. Manipulations of natural speech were allowed (e.g., low-pass filtered or resynthesised speech).
(vi) The data is not duplicated in the meta-analysis. In the case that the same data is represented in multiple eligible publications, the data from the first peer-reviewed publication were included.

Any response measures (e.g., behavioural, neurophysiological) and any test paradigms (e.g., visual fixation, head-turn preference) were considered. Any document with unique data was allowed regardless of publication status or type of publication, and documents from any years were considered.

2.2. Information sources and search strategy

As a first step, published studies ($n = 25$) and an unpublished dataset ($n = 1$) already known to the investigators were included. To search for additional records, the first author conducted a Google Scholar search on 17th April 2020 using Harzing's Publish or Perish Windows GUI Edition 7.19 software with the following keyword combination:

{"infant" OR "infancy" OR "baby"} & {"language discrimination" OR "dialect discrimination" OR "accent discrimination" OR "rhythm class discrimination"}.

This search yielded over 3000 records, the first 500 of which were considered, and three unique studies were retained after abstract and full text screening. To identify infant studies that focussed on durational metrics, we added the term 'deltaC' to the search term, as $\triangle$C was one of the first durational metrics examined in infant's language discrimination (Ramus et al., 1999) so we expect it to be mentioned in any subsequent studies that consider durational metric in language discrimination. Thus, the first author conducted a second Google Scholar search on 17th April 2020 with the keyword combination:

{"infant" OR "infancy" OR "baby"} & {"deltaC"} & {"rhythm"}.

This search yielded 299 results, all of which were considered, but no unique eligible studies were identified. Calls for studies were posted on the following mailing lists on April 16th, 2020: ICIS listserv, cogdevsoc listserv and the CHILDES mailing list, and three eligible studies were identified. The first author checked the reference lists of all included studies and one review on the topic (Nazzi & Ramus, 2003), revealing eight unique eligible studies. We requested recommendations for studies from corresponding authors of included studies who could be contacted, and two eligible studies were thus identified. The first author screened all titles for their eligibility based on the six criteria listed in Section 2.1. All duplicate titles were removed at this stage. Of the titles deemed

relevant, the first author screened all abstracts, and of all abstracts deemed relevant, she sought and scanned full articles to assess their eligibility. If an abstract could not be obtained, the full text was screened, and if a full text could not be obtained it was excluded. Fig. 1 shows a PRISMA-style flowchart of the identification and exclusion process.

In addition to the first author screening all titles and abstracts from the above sources, 10% of the abstracts that the first author deemed relevant from the title were double screened by the second or fourth author (34 of 318 eligible abstracts). An inter-rater reliability analysis using the R package *irr* (Gamer, Lemon, Fellows, & Singh, 2019) resulted in kappa = 0.674, which indicates moderate agreement. All disagreements bar one arose because the first author was more inclined to accept an abstract and check the full text, while the double screeners rejected the abstract; and in all these cases the first author's full text eligibility decision was also to reject the document. One other disagreement was resolved by discussion because the study did not fit the criterion of containing continuous speech (Phan & Houston, 2009), and so was excluded. One study that narrowly missed out on fulfilling the inclusion criteria was Paillereau et al. (2021) whose study does not include two different language varieties, but rather all Czech stimuli in a natural condition versus a temporally-manipulated condition. Pons, Bosch, and Lewkowicz (2015) investigated infants' gaze at a speaker's mouth and eyes when speaking Spanish versus Catalan, but we excluded this on the basis that the key component of the task was not discrimination or preference between two languages, and so including studies like this which are not designed to maximise differences based on discrimination abilities might add noise to the data. Five studies eligible for inclusion were not included in the meta-analysis because sufficient information could not be obtained to calculate effect sizes (see Section 2.4; Bosch, 2010; Diehl, Varga, Panneton, Burnham, & Kitamura, 2006; Johnson & Braun, 2011; Peña, Pittaluga, & Mehler, 2010; Sato et al., 2012), but are considered in the qualitative synthesis of results. All screened records, screening decisions and justifications for exclusion can be found in the *screening-decisions* spreadsheet (https://docs.google.com/spreadsheets/d/1ewIEhpx88ZIkvERmVjCMJ6milr-5Scqj3WzJU-yH4Kg/edit?usp=sharing) in our OSF repository.

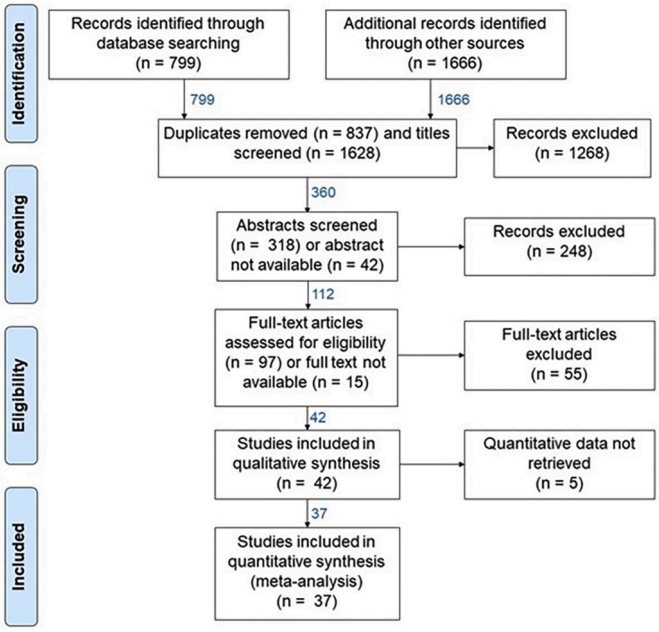


Fig. 1. PRISMA flowchart of identification, screening, eligibility, and inclusion processes.

2.3. Extraction of moderating factors

The critical variables for the planned analyses were the *task* (whether it was a discrimination or preference task), *method* (e.g., head-turn preference procedure, electroencephalography); *rhythm class* (whether the tested languages were in the same or different rhythm classes); *stimulus manipulation* (e.g., segmental, from low-pass filtering or resynthesis), *mean age* (in days) and durational metrics of the tested languages (see Table 1). All levels for each factor and the number of datapoints per factor level are reported in Tables 2 and 3. Factors used in exploratory analyses were *same language* (whether the tested languages were two distinct languages or two accents of the same language), *native language* (whether both, one or neither of the tested languages were native to the participants) and *language background* (monolingual/bilingual). Other relevant data were coded according to MetaLab guidelines (Bergmann et al., 2018, e.g., dependent measure, number of excluded participants; see OSF for full list and Code Book). Data were extracted from full texts and missing data were requested from authors by email.

Languages were assigned to rhythm class based on descriptions in the publication and/or the studies from which we extracted durational metrics (see Appendix B). For some languages, rhythm class categorisations were not wholly straightforward because of mixed results in the literature. In these cases, we chose the rhythm class to which a language was predominantly assigned in the literature (e.g., Tagalog, Catalan, Finnish as syllable-timed Bird, Fais, & Werker, 2005; Prieto, Vanrell, Astruc, Payne, & Post, 2012; White et al., 2016). As our analyses are intended to test the validity of the three traditional rhythm classes, we included these cases, and argue that tenuous assignment to rhythm class is a limitation of rhythm class itself. In some cases of non-standard varieties where we could not find information on the rhythm class, we assumed the variety was in the same rhythm class as the standard variety (e.g., Taiwanese Mandarin and Quebecois French were categorised as syllable-timed, and Scottish English as stress-timed). The only languages that we could not assign to a rhythm class were foreign-accented varieties (Chinese-accented English, English-accented French, French-

Table 2
Moderating factors included in rhythm class analysis meta-analytic model, levels, and contrast coding.

Factor	No. levels (k)	Levels	No. per level	Contrast coding
Task	2			Successive difference (baseline = grand mean)
		Preference	52	−0.5
		Discrimination	104	0.5
Rhythm class difference	2			Successive difference (baseline = grand mean)
		Same	89	−0.5
		Different	67	0.5
Stimulus manipulation	4	None	120	Simple (baseline = None)
		Segmental	31	−0.25
		Intonation	3	−0.25
		Segmental/ Intonation	2	−0.25
Method	4	HPP	67	Simple (baseline = HPP)
		HAS	23	−0.25
		CF	35	−0.25
		EEG/NIRS	31	−0.25
Factor		Scaling	Mean	SD
Mean age		z-scaled and centred	138.41 days	77.62 days

Table 3
Moderating factors included in durational metrics analysis meta-analytic model, levels, and contrast coding.

Factor	No. levels (k)	Levels	No. per level	Contrast coding
Task	2			Successive difference (baseline = grand mean)
		Preference	45	−0.5
		Discrimination	80	0.5
Method	4	HPP	66	Simple (baseline = HPP)
		HAS	18	−0.25
		CF	14	−0.25
		EEG/NIRS	27	−0.25
Factor		Scaling	Mean	SD
Mean age		z-scaled and centred	143.53 days	75.53 days
%V difference		z-scaled and centred	4.69	3.549
△C difference		z-scaled and centred	9.930	6.657
△V difference		z-scaled and centred	11.437	8.248
VarcoV difference		z-scaled and centred	6.656	5.837
nPVI-V difference		z-scaled and centred	11.935	9.029
rPVI-C difference		z-scaled and centred	10.280	6.902

accented English, and Spanish-accented English), which were excluded from analyses involving the factor rhythm class.

Durational metrics of any test languages that were not reported in an included study were obtained from other studies already known to the authors or found through Google Scholar searches, and weighted averages were calculated for each language variety, with weights defined as the number of sentences or utterances analysed in a study. "The North Wind and the Sun" is a commonly-used passage for cross-linguistic comparisons of auditory stimuli for phonetic analyses, which has 5 sentences, 113 words and around 44 s duration in English (Verhoeven, 2020), so when words or duration were reported instead of number of sentences, we estimated the number of sentences based on this conversion. If no information about the amount of data was reported we assumed that a minimum of 5 sentences were used, as no study reported using fewer than 5 sentences.

2.4. Effect size calculation and coding

The outcome of interest is the difference of response measures between test and control conditions, standardised as Hedges' *g*. Hedges' *g* is the ratio of the difference between the two conditions of interest over the pooled standard deviation, and scaled so that, in comparison to Cohen's *d*, data-points with smaller samples are shifted closer to 0 (Hedges, 1981). Means, standard deviations (SDs), *t*- and *F*-values, and sample size were extracted from studies or requested from authors, to calculate Hedges' *g* effect sizes and variance. If the authors could not provide the data-points, the values were deduced from figures where possible using WebPlotDigitizer software (https://apps.automeris.io/wpd/). For the effect sizes based on within-subject comparisons between responses to the two conditions, we additionally estimated the within-subject correlation between the means of the conditions. Exact correlation coefficients could be calculated for 51 records (median = 0.51); 97 were estimated from the means, SDs and *t*- or *F*-values, and correlations were imputed for the remaining 12 (from median = 0.563, variance = 0.471 with a floor/ceiling of ±0.961). Where possible, Hedges' *g* was calculated from means and SDs (Lipsey & Wilson, 2001), otherwise from *t*- or *F*-values and correlation coefficients (Dunlap, Cortina, Vaslow, & Burke, 1996). Effect sizes were calculated for 160 records from 90 experiments, including 2338 unique participants.

Data were coded so that a positive effect size indicates a novelty effect, and a negative effect size indicates a familiarity effect. In experiments with a pre-test exposure (habituation or familiarisation) phase, the familiar condition is defined as the language of exposure, and the novel condition as the new language encountered in the test phase. In experiments with no pre-test exposure phase, the familiar condition is defined as the native (or dominant, more familiar) language and the novel condition as the non-native language. For neurophysiological amplitude response measures, the absolute difference between conditions was calculated, and the sign of the effect size was coded based on which condition showed a greater absolute value. For example, if the native language yielded a greater response in the negative direction than the non-native language, the effect size would be coded as negative, indicating a familiarity effect. For all records, a positive effect size indicates that the infants responded more strongly (looked longer, oriented fasted or exhibited stronger or earlier brain activity) to the non-familiarised, non-habituated or non-native language, depending on the experiment design. Meanwhile, a negative effect size indicates a stronger response to the native, familiarised, or habituated variety.

2.5. Synthesis of results

Quantitative analysis was conducted in R and RStudio (R Core Team, 2020; RStudio Team, 2020), with code adapted from MetaLab (http://metalab.stanford.edu) and previous publications using the MetaLab framework (Black & Bergmann, 2017; Carbajal, Peperkamp, & Tsuji, 2021; Rabagliati, Ferguson, & Lew-Williams, 2019). Inverse-variance-weighted random effects multivariate meta-analytic regression models were created using the *metafor* package (Viechtbauer, 2010), plots were created using *ggplot2* (Wickham, 2016), and the *MASS* package was used for contrast coding (Venables & Ripley, 2002).

2.5.1. Baseline analysis: discrimination vs. preference tasks

A model was run to calculate an overall estimated effect size, by entering vectors of Hedges' *g* and its variance. Sometimes the same participants contributed data to more than one experiment within a study. Relatedly, we expect some shared covariance from results in the same experiment or the same study. To account for covariance between results from the same experiment, participants, or publication, we added random effects nested experiment in participant, and participant in study. Discrimination and preference tasks are not expected to necessarily show effects in the same direction, since familiarity with and exposure to stimuli are considered to influence the direction of effects. Discrimination tasks that have a pre-test habituation or familiarisation phase have most commonly been found to show novelty effects, while the direction of preference tasks are less reliable, but tend to yield familiarity effects with continuous, natural language stimuli (see Bergmann, Rabagliati, & Tsuji, 2019; Houston-Price & Nakai, 2004; Oakes, 2010 for discussions). Due to possible differences in the directions of effects, the same model was run with *task* added as a moderator to show the estimated difference in discrimination and preference task effect sizes.

2.5.2. Rhythm class analysis

Four records were excluded from the rhythm class analysis, because the tested languages could not be classified as belonging to a rhythm class: Chinese-accented English (Chung, 2002), English-accented French and French-accented English (Kinzler, Dupoux, & Spelke, 2007; White et al., 2014), Spanish-accented English (Paquette-Smith & Johnson, 2015). A model was built with the following five moderating variables: as numerator: (1) *task*; as denominator: the three-way interaction of (2) *rhythm class difference*, (3) *mean age* (in days) and (4) *stimulus manipulation*; and (5) *method*. Placing *task* as the numerator allows one to

observe the moderating factors in the denominator separately by discrimination and preference, because effect sizes are expected to have opposite polarities on average. The method Forced Choice (FC) was only represented in two data-points (both preference tasks), and due to convergence issues, it was conflated with the Central Fixation (CF) method. Table 2 lists all moderators, their levels and respective contrast codings.

We conducted model comparisons of the maximal model to a model with one factor of interest at a time excluded. The effects of any factors whose inclusion lowered the model's Akaike Information Criterion (AIC; Akaike, 1974) were included in what we consider the "best-fitting" model. However, the results of the Likelihood Ratio Tests (LRT) comparing the reduced and full model were taken as the indication of whether a factor's inclusion in the model as a whole was significant at the level of $p < 0.05$. The AIC and LRT do not, however, indicate whether a factor's effect is significant in discrimination or preference tasks, or both. To determine this, as well as the size of significant effects, we then appraised the effects of all factors included in the best-fitting model in the model output.

2.5.3. Analysis of durational metrics

We sought all durational metrics that had been examined by White et al. (2014) and White et al. (2012) in discrimination and preference studies, but decided to only include durational metrics in the analyses if they could be obtained for over 75% of included language varieties. Six durational metrics reached this threshold: %V, △C, △V, VarcoV, nPVI-V, rPVI-C and so were included in the main analysis. See Appendix B for these metrics in all included language varieties, and plotted by rhythm class as they were in previous studies (Grabe & Low, 2002; Nespor, Shukla, & Mehler, 2011; Ramus et al., 1999; White & Mattys, 2007).

For each of the six durational metrics for each record, the difference in that metric between the two language varieties tested for discrimination or preference was calculated. Thirty-five records were excluded due to missing values for Miami English and Cuban Spanish (Bahrick & Pickens, 1988), Tagalog (Byers-Heinlein et al., 2010; May, Byers-Heinlein, Gervain, & Werker, 2011), New York Hispanic English and Taiwanese Mandarin (Chung, 2002), Quebecois French (Cristia, 2013), English-accented French (Kinzler et al., 2007), South African English (Kitamura et al., 2013), Basque (Molnar et al., 2013; Molnar & Carreiras, 2015), and Western Catalan (Zacharaki & Sebastián-Gallés, 2019). Since stimulus manipulation failed to show any significant effects in the first analyses (see Section 3.2), it was not included in subsequent analyses.

Including multiple continuous factors and interaction terms in a model can introduce collinearity which can reduce the interpretability and stability of estimates (Afshartous & Preston, 2011; Aiken, West, & Reno, 1991). Thus, the differences in the six durational metrics and their interaction with mean age (all centred and z-transformed, see means and standard deviations in Table 3) were inspected for collinearity using variance inflation factors (VIF, see White et al., 2014). The interaction of △V and mean age had the highest VIF over the threshold of five, and so was excluded, which resulted in all remaining VIFs being less than five. These remaining factors and interactions were added to the denominator of a multivariate random effects meta-analytic model, along with method, with task as the numerator, and with the same random effects as previous models. As previously, one factor was removed at a time and compared to the maximal model by inspecting the AIC for better model fit and using the Likelihood Ratio Test.

2.5.4. Comparison of rhythm class and durational metrics

The factor rhythm class and its interaction with age were added to the best-fitting durational metrics model (see Section 3.3), and further model comparisons were conducted as above, by removing one factor at a time from the full model and inspecting the AIC and LRT. This allowed investigation of the extent to which durational metrics accounted for the same variance as the factor rhythm class in discrimination and preference effect sizes.

2.5.5. Exploratory analyses with additional moderators

In exploratory analyses, the following factors were added to the model identified as best-fitting from the previous analysis (see Section 3.4) to determine their effects: *same language* (whether the experiment tested two distinct languages or two accents of the same language), *native language* (whether both, one, or neither of the tested languages were native to the participant), *language background* (monolingual/bilingual).

3. Results

This section presents the results of the analyses described in Section 2.5. First, the overall effect sizes with and without the effect of *task* are presented (3.1), followed by the results of the rhythm class analysis (3.2) in discrimination tasks (3.2.1) and in preference tasks (3.2.2). Then, the results of the durational metrics analysis are provided (3.3) in discrimination tasks (3.3.1, no significant effects were found in preference tasks in this analysis). Next, the comparison of rhythm class and durational metrics (3.4) and the results of exploratory analyses (3.5) are presented. The risk of bias (3.6) in discrimination (3.6.1) and preference tasks (3.6.2) is considered, followed by a summary of results (3.7).

3.1. Baseline analysis: discrimination vs. preference tasks

See Appendix C for study characteristics. A forest plot can be found in Appendix D showing all discrimination effect sizes, and in Appendix E showing all preference effect sizes. These illustrate all calculated effect sizes and their 95% confidence intervals (CIs), in order of magnitude.

A random effects meta-analytic model with no moderating variables (see Section 2.5.1) estimated an overall effect size of $\widehat{\beta} = 0.068$ ($SE = 0.065$, $z = 1.054$, $p = 0.292$, see Appendix F for full model output), which is not significantly different from zero. When we added the moderating variable of *task* to separate discrimination and preference records, the effect of task was significant ($\widehat{\beta} = 0.513$, $SE = 0.094$, $z = 5.493$, $p < 0.0001$, where $\widehat{\beta}$ is the estimated difference between tasks, see Appendix G for full model output). This justifies investigating records separately by task. The overall estimated effect size for discrimination tasks is $\widehat{\beta} = 0.254$ (95% CI: [0.162, 0.345]). The overall estimated effect size for preference tasks is $\widehat{\beta} = -0.259$ (95% CI: [−0.351, −0.168]). Thus, as hypothesised, these two types of task diverge in the overall direction of effect sizes such that discrimination tasks show a novelty preference, and preference tasks show a familiarity preference. Cochran's Q-test indicated significant residual heterogeneity ($Q(158) = 723.361$, $p < 0.0001$), which justifies investigating additional moderators.

3.2. Rhythm class analysis

The moderating variables of (i) rhythm class difference, (ii) mean age, (iii) stimulus manipulation and their interactions, and (iv) method, were entered into the model, nested by task (so that discrimination and preference effect sizes are examined separately, see Section 2.5.2). The effect of method was significant ($LRT = 13.218$, $p = 0.040$). Rhythm class showed a significant interaction with age ($LRT = 7.312$, $p = 0.026$, see Sections 3.2.1 and 3.2.2 for magnitude of effects in discrimination and preference tasks respectively). All other effects were non-significant ($p > 0.05$). The main effect of rhythm class decreased the AIC to yield a better model fit, however the difference between models was non-significant ($LRT = 4.516$, $p = 0.105$). The best-fitting model is displayed in Table 4 and includes the main effects of method and rhythm class, and the interaction of rhythm class and age (see Appendix H for full model output). Inspection of the model output indicates whether the factors affected discrimination or preference tasks, which will be described in turn in the following sections.

We also conducted a sensitivity analysis removing all effect sizes >2

Table 4
Results of rhythm class analysis meta-regression. Intercept reflects estimated grand mean. The effect of Task shows the estimated difference between discrimination and preference tasks, and effects of subsequent moderators are provided for Discrimination then Preference tasks.

		Estimate	SE	95% CI	z	p
1	Intercept	−0.0381	0.0635	[−0.1626, 0.0864]	−0.6000	0.5485
	Task					
2	Discrimination-Preference	0.5616	0.1239	[0.3187, 0.8045]	4.5316	**<0.0001****
	Discrimination					
3	Rhythm class (Different-Same)	0.2123	0.1012	[0.0139, 0.4108]	2.0970	**0.0360**
4	Same rhythm class: Mean age	−0.0399	0.0848	[−0.2061, 0.1262]	−0.4711	0.6376
5	Different rhythm class: Mean age	0.0693	0.1602	[−0.2446, 0.3832]	0.4324	0.6654
6	Method (HAS)	0.1269	0.2532	[−0.3693, 0.6231]	0.5013	0.6162
7	Method (CF)	0.3141	0.1293	[0.0607, 0.5675]	2.4291	**0.0151***
8	Method (EEG/NIRS)	−0.5328	0.2220	[−0.9679, −0.0976]	−2.3996	**0.0164***
	Preference					
9	Rhythm class (Different-Same)	−0.0614	0.1547	[−0.3646, 0.2418]	−0.3971	0.6913
10	Same rhythm class: Mean age	0.3426	0.1538	[0.0412, 0.6440]	2.2277	**0.0259***
11	Different rhythm class: Mean age	−0.1344	0.0879	[−0.3067, 0.0378]	−1.5296	0.1261
12	Method (HAS)	−0.0139	0.2921	[−0.5865, 0.5587]	−0.0476	0.9621
13	Method (CF)	−0.0344	0.1958	[−0.4181, 0.3493]	−0.1758	0.8605
14	Method (EEG/NIRS)	0.1356	0.2110	[−0.2781, 0.5492]	0.6424	0.5206

*LRT $p < .05$, **LRT $p < .01$, bold indicates model estimate $p < 0.05$.

SD from the mean. Results showed that the main effect of rhythm class did not lower the model AIC. This suggests that the effect of rhythm class may be spurious and sensitive to outliers. All other significant effects and best-fitting models presented in this paper were not changed by inclusion or exclusion of outliers.

3.2.1. Rhythm class analysis in discrimination tasks

The effect of rhythm class was estimated by the model such that in discrimination tasks, languages tested in different rhythm classes will on average show a stronger novelty effect (more positive effect sizes) than those in the same rhythm class (Table 4, line 3: $\widehat{\beta} = 0.212$, $SE = 0.101$, $z = 2.097$, $p = 0.036$, but not significant according to the LRT). The factor of rhythm class did not show a significant interaction with age in discrimination tasks (line 4, same rhythm class: $z = -0.471$, $p = 0.638$; line 5, different rhythm class: $z = 0.432$, $p = 0.665$). Fig. 2 plots discrimination effect sizes by rhythm class.

The model estimates indicate that in discrimination tasks, the CF method yields significantly larger effect sizes than the baseline HPP (line 7: $\widehat{\beta} = 0.314$, $SE = 0.129$, $z = 2.429$, $p = 0.015$). The model estimated that neurophysiological (EEG/NIRS) studies using discrimination tasks yield on average more negative effect sizes than HPP (line 8: $\widehat{\beta} = -0.533$, $SE = 0.222$, $z = -2.400$, $p = 0.016$). Fig. 3 illustrates how EEG/NIRS discrimination measures tend to yield both novelty (positive) and familiarity (negative) effects, while behavioural methods more reliably yield novelty effects.

Due to a particular interest in the literature regarding the discrimination abilities of newborns, we conducted an exploratory subset analysis of discrimination effect sizes in newborns only. These effect sizes are plotted in Fig. 4. This analysis brought to light the fact that only one record tested newborns' ability to discriminate between two languages in the same rhythm class (Nazzi et al., 1998; line 1 of Fig. 4, $g = 0.25$, 95% CI [−0.04, 0.54]). Meanwhile the estimated overall effect size of newborns' discrimination of rhythmically distinct languages was estimated to be significantly different from zero ($\widehat{\beta} = 0.338$, 95% CI: [0.202, 0.474], from Byers-Heinlein et al., 2010; Mehler et al., 1988; Nazzi et al., 1998; Ramus, 2002; Ramus, Hauser, Miller, Morris, & Mehler, 2000, see line 2 onwards of Fig. 4).

3.2.2. Rhythm class analysis in preference tasks

The best-fitting model showed that there was no significant main

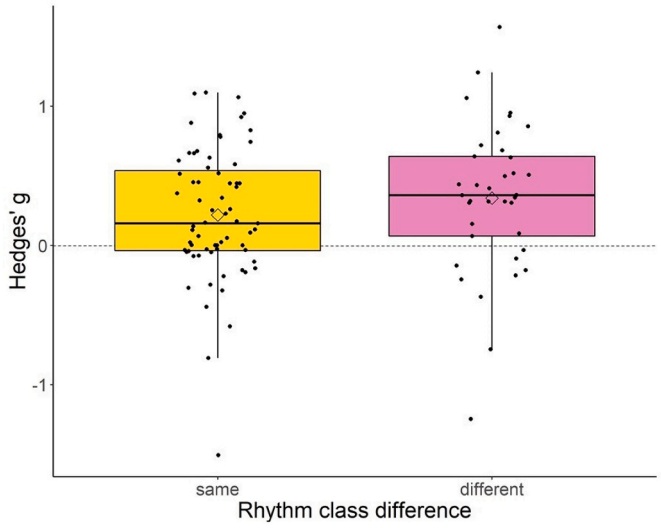


Fig. 2. Boxplot of discrimination effect sizes (y-axis) by rhythm class difference (x-axis: same or different). Diamonds indicate mean values.

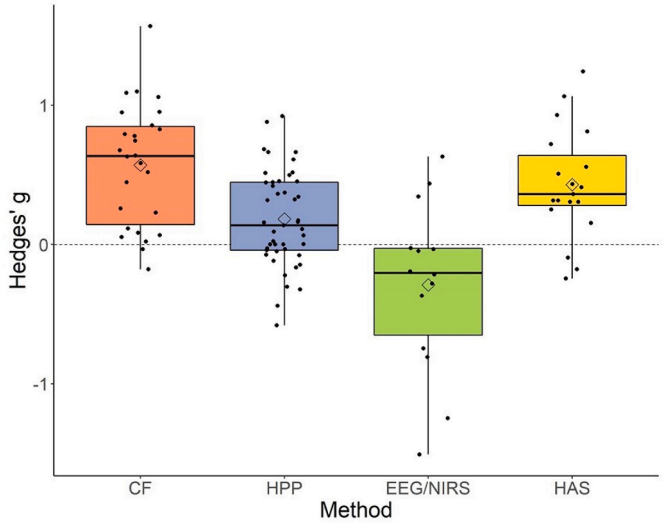


Fig. 3. Boxplot of discrimination effect sizes (y-axis) by method (x-axis; CF = central fixation, HPP = head-turn preference procedure, EEG/NIRS = neurophysiological, HAS = high-amplitude sucking). Diamonds indicate mean values.

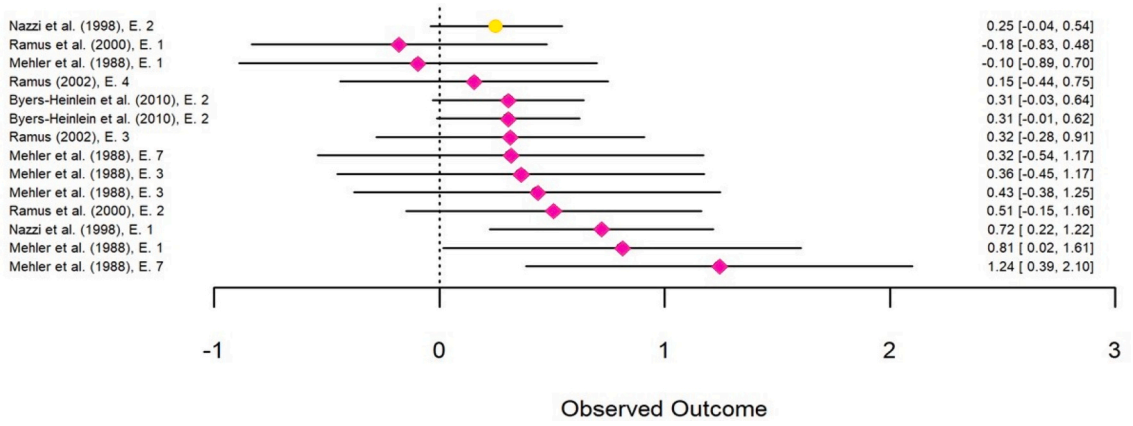


Fig. 4. Forest plot of discrimination effect sizes and 95% CIs in newborns (x-axis, and y-axis, right) by study experiment (y-axis, left). Colour- and shape-coded by rhythm class difference (same in yellow/circle, different in pink/diamond). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

effect of rhythm class in preference tasks (Table 4, line 9: $z = -0.397$, $p = 0.691$). There was, however, a significant interaction of rhythm class and age, which was estimated by the model such that in preference tasks that test languages (native versus non-native) in the same rhythm class, effect sizes become more positive (indicating a stronger preference for the non-native language) with age (line 10, $\widehat{\beta} = 0.343$, $SE = 0.154$, $z = 2.228$, $p = 0.026$). Specifically, the difference in effect sizes between different and same rhythm class comparisons is estimated to increase by 0.343 as age increases by 2.54 months. This indicates a growing preference with age for a non-native language variety that is in the infant's native rhythm class. This is not the case when the non-native language is in a non-native rhythm class, where the model estimates an increasing native language preference with age (signified by a negative slope), however this is non-significant (line 11, $\widehat{\beta} = -0.134$, $SE = 0.088$, $z = -1.530$, $p = 0.126$). This effect is illustrated in Fig. 5.

3.2.3. Summary of rhythm class analysis

Rhythm class accounted for infants' discrimination performance in that effect sizes were larger when the tested languages were in different rhythm classes. In preference tasks, there was an increasing preference with age for the non-native language when it was in the native rhythm class, not when they were in a non-native rhythm class. In discrimination tasks, the CF method yielded the largest effect sizes, and the EEG/NIRS methods tended to yield both familiarity and novelty effects. Stimulus manipulation failed to show any significant effects or interactions. This model showed significant residual heterogeneity (Q (142) = 623.038, $p < 0.0001$), which indicates there is substantial variance in the data not accounted for by the included moderators. The following analysis will investigate the role of durational metrics instead of rhythm class in accounting for discrimination and preference.

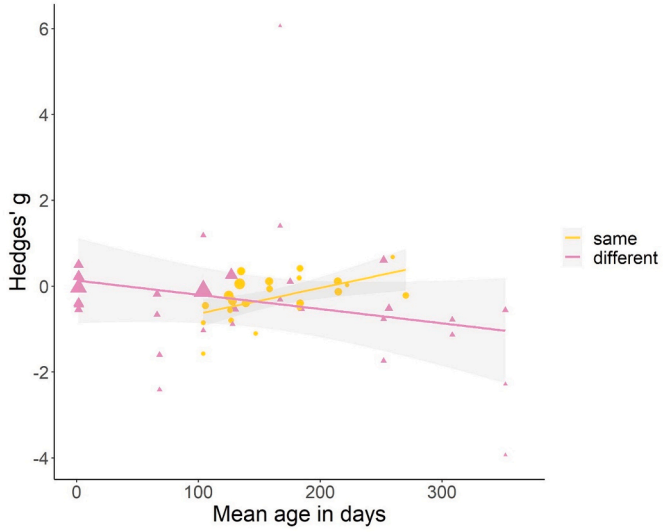


Fig. 5. Preference effect sizes (y-axis; negative values indicate native language preference) by age in days (x-axis). Colour- and shape-coded by rhythm class (same: native and non-native language from same rhythm class; different: native and non-native language from different rhythm classes) and weighted by inverse of standard error.

3.3. Analysis of durational metrics

The moderating variables of (i) method, (ii) age, (iii) %V, (iv) △C, (v) △V, (vi) VarcoV, (vii) nPVI-V, (viii) rPVI-C, and the interactions of each durational metric except for △V with age were entered into the model, nested by task (see 2.5.3). Full model comparisons revealed a main effect of △V ($LRT = 6.921$, $p = 0.031$) and a main effect of rPVI-C ($LRT = 8.423$, $p = 0.015$). All other factors and interactions did not lower the AIC and were non-significant ($p > 0.05$). This includes the effect of method, which, unlike in the previous analyses, was non-significant ($LRT = 11.133$, $p = 0.084$). Thus, the best-fitting model included only △V and rPVI-C, nested in task (see model output in Table 5, Appendix I).

No effects were significant in preference tasks (lines 5 and 6), so the following section discusses the effects of △V and rPVI-C in discrimination tasks. In an exploratory analysis, minimal models were run with each durational model separately to investigate their individual effects. See Appendix J for output of these models.

3.3.1. Analysis of durational metrics in discrimination tasks

The model output revealed that the factor △V was significant in discrimination tasks (Table 5, line 3: $\widehat{\beta} = -0.173$, $SE = 0.059$, $z = -2.945$, $p = 0.003$), whereby effect sizes show a weaker novelty effect (i.e. become more negative) the more the two tested languages differ in △V. rPVI-C was significant in discrimination tasks (line 4: $\widehat{\beta} = 0.148$, $SE = 0.052$, $z = 2.838$, $p = 0.005$), whereby effect sizes show a stronger novelty effect (became more positive) the more that the two tested languages differ in rPVI-C. Discrimination effect sizes are plotted by △V

Table 5
Results of durational metrics analysis meta-regression. Intercept reflects estimated grand mean when differences in △V and rPVI-C are 11.437 and 10.280, respectively. The effect of Task shows the estimated difference between discrimination and preference tasks, and effects of subsequent moderators are provided for Discrimination then Preference tasks.

		Estimate	SE	95% CI	z	p
1	Intercept	−0.0494	0.0756	[−0.1976, 0.0988]	−0.6532	0.5136
	Task					
2	Preference-Discrimination	0.5151	0.1509	[0.2195, 0.8108]	3.4147	**0.0006****
	Discrimination					
3	△V	−0.1726	0.0586	[−0.2875, −0.0577]	−2.9446	**0.0032***
4	rPVI-C	0.1478	0.0521	[0.0457, 0.2498]	2.8382	**0.0045***
	Preference					
5	△V	0.0233	0.0778	[−0.1293, 0.1758]	0.2991	0.7649
6	rPVI-C	−0.0005	0.1060	[−0.2083, 0.2073]	−0.0049	0.9961

*LRT $p < .05$, **LRT $p < .01$, bold indicates model estimate $p < 0.05$.

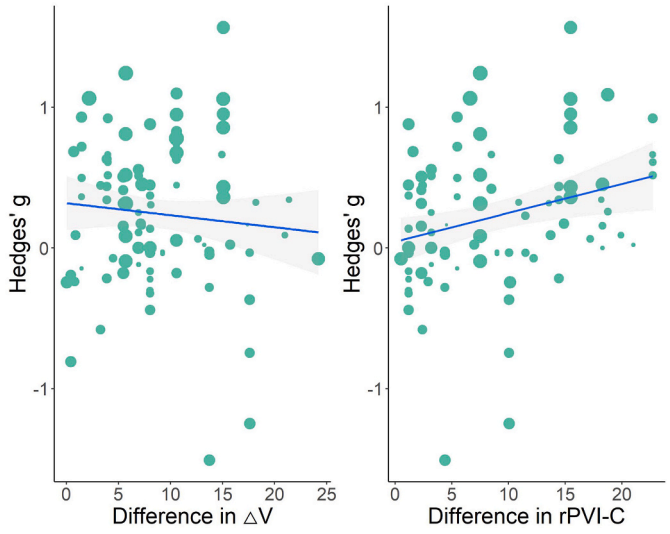


Fig. 6. Discrimination effect sizes (y-axis) by difference in △V (left, x-axis) and rPVI-C (right, x-axis) between the tested languages.

and rPVI-C in Fig. 6.[2]

3.4. Comparison of rhythm class and durational metrics

Model comparisons revealed that the best-fitting model according to the AIC included the main effect of rhythm class, interaction of rhythm class and age, △V, and rPVI-C. The effects of the two durational metrics remained significant (△V: $LRT = 9.743$, $p = 0.008$, rPVI-C: $LRT = 8.173$, $p = 0.017$) when these additional factors were included in the model. Inspection of the model output revealed that with the inclusion of △V and rPVI-C, the factor of rhythm class mainly accounted for variance in preference tasks (Table 6, line 10, $\widehat{\beta} = -0.451$, $SE = 0.269$, $z = -1.678$, $p = 0.093$) and had a negligible effect in discrimination tasks (line 5: $\widehat{\beta} = 0.047$, $SE = 0.118$, $z = 0.399$, $p = 0.690$). The inclusion of the interaction of rhythm class and age also improved model fit, accounting for variance in preference tasks, as was seen in the previous rhythm class analysis (line 11). The relationships between rhythm classes, △V and rPVI-C are plotted in Figs. 7 and 8, showing that languages in the same rhythm class tend to cluster along these two metrics (compare with Figs. B.1–B.5 in Appendix B), but the distinction is not so clear when defined by a *difference* in rhythm class or in the durational metrics. The best-fitting model (Table 6, see Appendix K for full model output) showed significant residual heterogeneity ($Q(113) = 514.483$, $p < 0.0001$), which indicates there is still substantial variance in the data not accounted for by the included moderators.

3.5. Exploratory analyses with additional moderators

The above analyses showed that the best-fitting model included main effects of △V, rPVI-C, rhythm class, and the interaction of rhythm class and age, nested in task. The subsequent analyses compared this model with models including effects of *same language, native language,* and *language background*. These analyses showed no significant effects of *same language* or *native language*.

The effect of *language background* (monolingual/bilingual) was significant ($LRT = 9.311$, $p = 0.010$). The interaction of rPVI-C and language background was non-significant ($LRT = 5.318$, $p = 0.070$) but lowered the AIC. With the inclusion of these effects, the main effect of rhythm class no longer improved model fit. The output of the best-fitting model (see Appendix L) revealed that monolingual infants showed a significantly larger novelty effect in discrimination tasks than bilingual infants ($\widehat{\beta} = 0.522$, $SE = 0.246$, $z = 2.126$, $p = 0.034$, see Fig. 9), and a significantly larger familiarity effect in preference tasks than bilingual infants ($\widehat{\beta} = -0.831$, $SE = 0.220$, $z = -3.773$, $p = 0.0002$, see Fig. 10). The effect of rPVI-C on discrimination effect sizes was stronger in bilingual infants than in monolingual infants ($\widehat{\beta} = -0.327$, $SE = 0.147$, $z = -2.219$, $p = 0.027$, see Fig. 11). In all cases, note the small number of datapoints for bilingual infants compared to monolingual infants.

3.6. Risk of bias

Evidence for risk of bias should be considered when appraising the above results. Below, funnel plots show each data-point's effects sizes along the x-axis by its standard error (SE; an approximation of precision) on the y-axis, with data-points colour-coded by method and peer-reviewed status, and a funnel in white spreading ±1.96 *SE* centring around the calculated overall effect sizes to illustrate a 95% CI. In the case of no sample heterogeneity or publication bias, data-points should cluster within the 95% CI centred around the overall estimated mean (Sterne et al., 2011).

3.6.1. Risk of bias in discrimination tasks

The funnel plot in Fig. 12 shows discrimination effect sizes. Visual inspection reveals that many data-points sit outside of the 95% CI, which likely indicates sample heterogeneity. Data-points with large standard errors tend to have large positive effect sizes, which could be erroneously increasing the estimated effect size and non-peer-reviewed studies in general have smaller effect sizes and smaller SEs, both of which suggest the presence of publication bias. Studies using CF and High-

[2] Analyses removing all effect sizes >2 SD from the mean did not change this result, just as all other significant effects and best-fitting models presented in this paper were not changed by inclusion or exclusion of outliers (except for the effect of rhythm class in the Rhythm class analysis, Section 3.2).

Table 6
Results of durational metrics and rhythm class meta-regression. The effect of Task shows the estimated difference between discrimination and preference tasks, and effects of subsequent moderators are provided for Discrimination then Preference tasks.

		Estimate	SE	95% CI	z	p
1	Intercept	−0.1047	0.0766	[−0.2548, 0.0454]	−1.3669	0.1717
	Task					
2	Discrimination-Preference	0.6568	0.1533	[0.3564, 0.9572]	4.2857	**<0.0001****
	Discrimination					
3	△V	−0.1511	0.0608	[−0.2703, −0.0319]	−2.4843	**0.0130****
4	rPVI-C	0.1499	0.0519	[0.0482, 0.2515]	2.8899	**0.0039***
5	Rhythm class	0.0469	0.1176	[−0.1835, 0.2773]	0.3989	0.6900
6	Same rhythm class: Mean age	−0.0700	0.1088	[−0.2832, 0.1431]	−0.6438	0.5197
7	Different rhythm class: Mean age	−0.1110	0.1122	[−0.3308 0.1088]	−0.9895	0.3224
	Preference					
8	△V	0.1767	0.0989	[−0.0171, 0.3705]	1.7866	0.0740
9	rPVI-C	−0.0105	0.1141	[−0.2342, 0.2131]	−0.0924	0.9263
10	Rhythm class	−0.4509	0.2686	[−0.9774, 0.0757]	−1.6783	0.0933
11	Same rhythm class: Mean age	0.4832	0.2044	[0.0827, 0.8838]	2.3646	**0.0180***
12	Different rhythm class: Mean age	−0.0637	0.0883	[−0.2367, 0.1094]	−0.7208	0.4710

*LRT p < .05, **LRT p < .01, bold indicates model estimate p < 0.05.

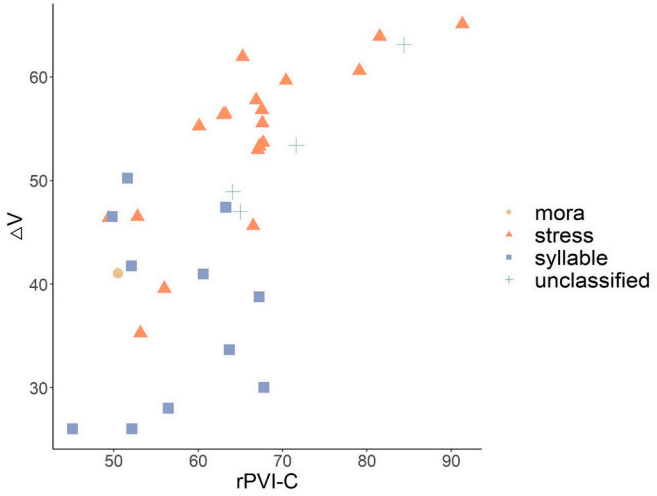


Fig. 7. △V (y-axis) and rPVI-C (x-axis) of language varieties. Colour- and shape-coded by rhythm class (mora, stress, syllable, unclassified).

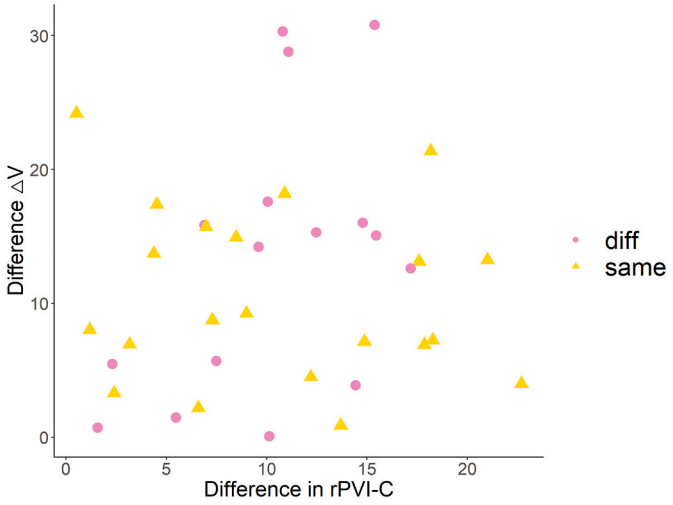


Fig. 8. Difference in △V (y-axis) and in rPVI-C (x-axis) of language varieties. Colour- and shape-coded by rhythm class difference (different or same).

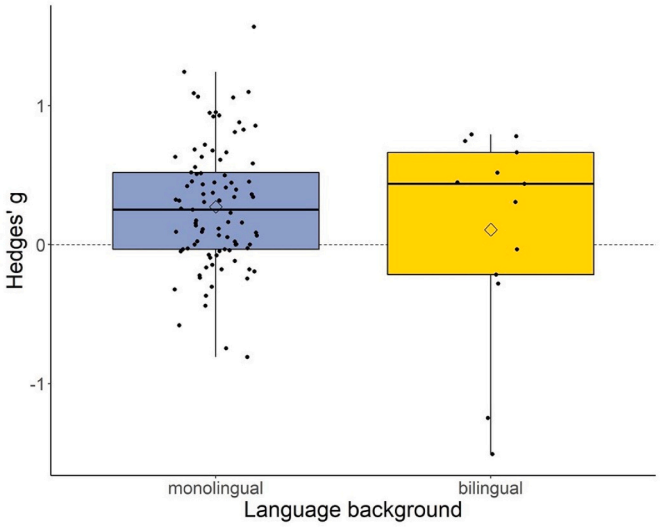


Fig. 9. Boxplot of discrimination effect sizes (y-axis) by language background (x-axis; monolingual or bilingual). Diamonds indicate mean values.

Amplitude Sucking (HAS) methods tended to yield both large effect sizes and large standard errors, while HPP studies yielded effect sizes and SEs closer to zero. As was seen in 3.2.1 (Fig. 3), EEG/NIRS studies tended to have more negative effect sizes.

Egger's test for funnel plot asymmetry confirms significant asymmetry in peer-reviewed studies ($z = 2.525$, $p = 0.012$) and non-peer-reviewed studies ($z = 2.290$, $p = 0.022$), which suggests bias. There was, however, a low correlation of sample size (n) and effect size (g) (Kendall's tau = −0.107, $p = 0.118$), which indicates that large effect sizes do not arise only in the case of small sample size.

3.6.2. Risk of bias in preference tasks

Funnel plots are shown in Fig. 13 (in the plot on the right the three outlying data-points are removed for visual clarity of smaller data-points). Egger's test for funnel plot asymmetry did not show significant asymmetry in peer-reviewed ($z = -0.835$, $p = 0.404$) or non-peer-reviewed ($z = 1.203$, $p = 0.229$) preference tasks, and a low correlation of n and g suggests little evidence for publication bias (Kendall's tau = 0.079, $p = 0.425$). However, visual inspection of the funnel plots again reveals that data-points do not cluster within the 95% CI, and data-points with larger SEs tend to be negative, which may be pulling the overall effect size in this direction. This could suggest sample

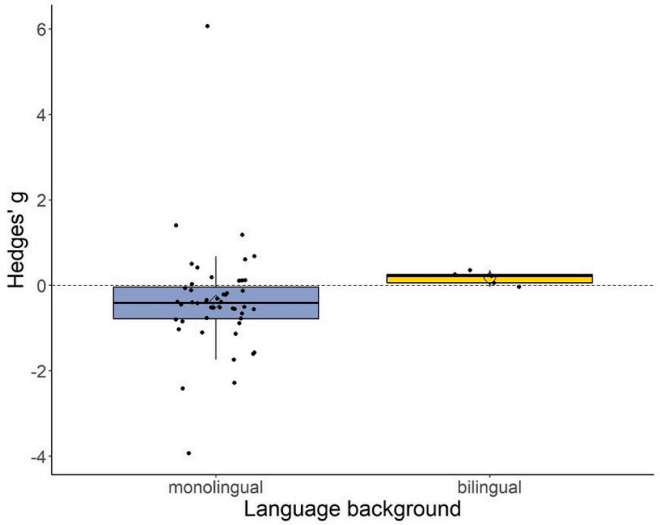


Fig. 10. Boxplot of preference effect sizes (y-axis) by language background (x-axis; monolingual or bilingual). Diamonds indicate mean values.

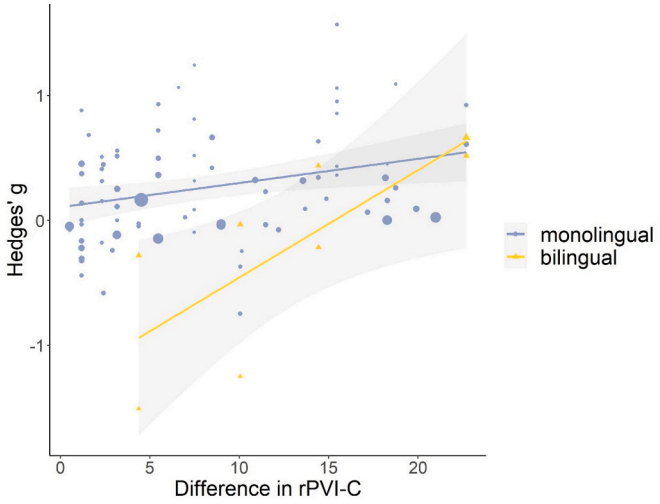


Fig. 11. Discrimination effect sizes (y-axis) by difference in rPVI-C (x-axis). Colour- and shape-coded by language background (monolingual or bilingual) and weighted by inverse standard error.

heterogeneity, reporting bias or methodological artifacts.

3.7. Summary of results

The results showed that in discrimination tasks, the main effect of rhythm class accounted for variance in effect sizes, whereby effect sizes were larger when languages were in different rhythm classes. This effect was not significant according to the LRT, however, and subsequent analyses showed that differences in discrimination effect sizes are better accounted for by the durational metrics △V and rPVI-C. This finding was such that discrimination effect sizes became larger as differences in △V decreased, and differences in rPVI-C increased.

In preference tasks, there was a significant interaction of rhythm class and age, whereby if languages were in the same rhythm class, there was an increasing non-native preference with age. This was not the case for languages in different rhythm classes. No durational metrics suitably accounted for this finding.

Bilingual infants showed less of a familiarity effect in preference tasks and less of a novelty effect in discrimination tasks than monolinguals. The effect of rPVI-C on discrimination effect sizes was more pronounced in bilinguals. In discrimination records in the rhythm class analysis, the CF method yielded larger effect sizes, and EEG/NIRS yielded more negative effect sizes than behavioural methods. This effect did not remain significant in the durational metrics analysis. A summary of findings is presented in Table 7.

4. Discussion

In this meta-analysis two research questions were posed:

1) How do typically-developing infants' ability to discriminate between languages in the same or different rhythm classes change from birth up to 12 months of age?
2) Which durational cue(s) best predict infants' language discrimination skills from newborns up to 12 months of age?

In Section 4.1, the first research question is addressed, by presenting a qualitative overview of the literature, synthesising this with the meta-analytic evidence obtained in Section 3. This sheds light on where results accord with or contradict the *rhythmic class acquisition* and *native language acquisition hypotheses* presented by Nazzi and Ramus (2003), and where evidence is still lacking. In Section 4.2, the second research question is addressed, discussing the finding that large differences in consonant interval variability and small differences in vowel interval variability better account for discrimination effect sizes than the factor rhythm class. It is discussed how these findings fit with Mehler and colleagues' (1996; Nespor et al., 2003; Ramus et al., 1999) prediction that infants would pay particular attention to vowel intervals. The

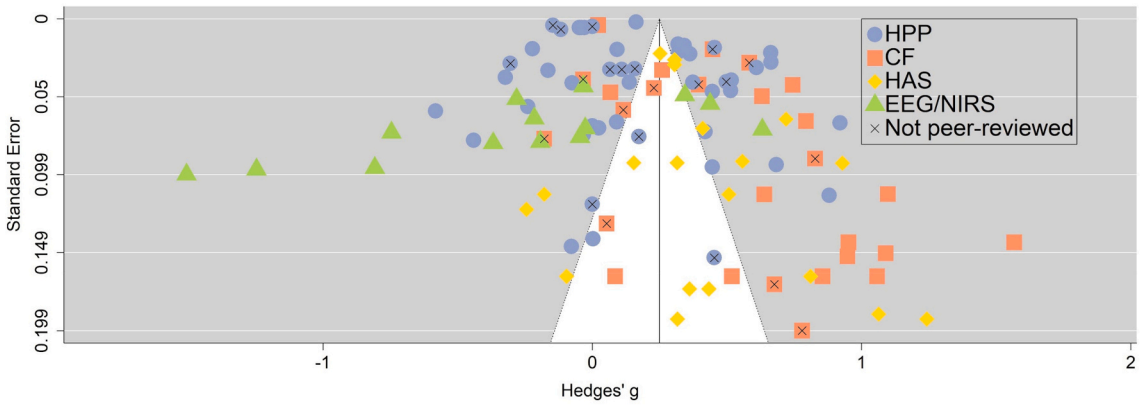


Fig. 12. Funnel plot of discrimination effect sizes. Colour- and shape-coded by method, and cross indicates a record was not peer-reviewed.

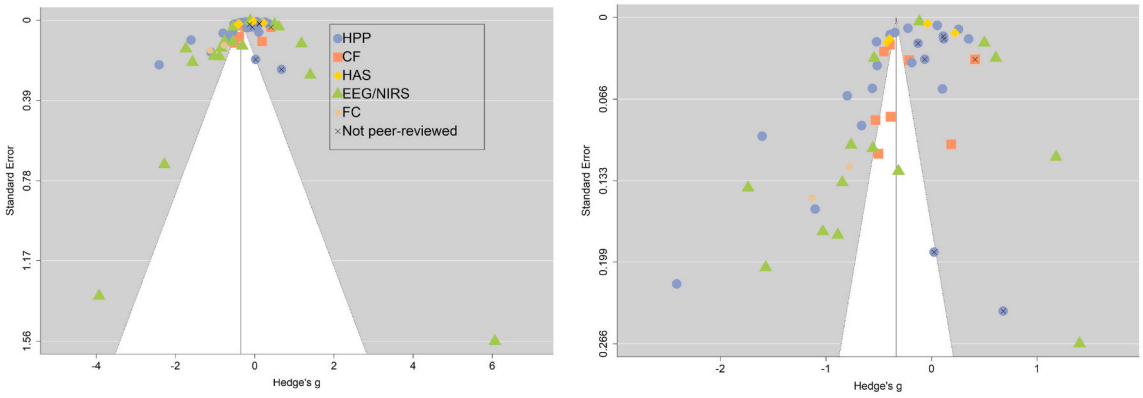


Fig. 13. Funnel plots of preference effect sizes: all values (left), outliers removed (right). Colour-coded by method, and outline indicates a record was not peer-reviewed.

Table 7
Summary of findings, separated by discrimination and preference tasks.

	Discrimination	Preference
Rhythm class	Different>Same (non-sig in later analyses)	–
Method (CF)	CF > HPP (non-sig in later analyses)	–
Method (EEG/NIRS)	EEG/NIRS<HPP (non-sig in later analyses)	–
Same rhythm class: Age	–	*g* increases as age increases
Different rhythm class: Age	–	–
△V	*g* decreases as difference in △V increases	–
rPVI-C	*g* increases as difference in rPVI-C increases	–
Language background	Smaller novelty effect in bilinguals	Smaller familiarity effect in bilinguals
Language background: rPVI-C	Effect of difference in rPVI-C larger in bilinguals	–

limitations and strengths of the study (Section 4.3) and areas for future research (Section 4.4) are then considered.

It was appropriate to analyse discrimination and preference tasks separately for the quantitative analysis because the predicted direction of effect sizes is opposite in each task type. However, our primary interest is how both types of task are informative regarding infants' discrimination skills. Therefore, unless otherwise specified, results of both discrimination and preference tasks are discussed in synthesis in terms of what they reveal about language discrimination.

4.1. Language discrimination over the first year of life

For the first research question, we asked how typically-developing infants' language discrimination skills change from birth up to 12 months of age. We predicted that discrimination effect sizes would be larger in younger infants for any between-rhythm-class contrasts compared to within-rhythm-class contrasts, based on the *rhythmic class acquisition* and *native language acquisition hypotheses* (Nazzi & Ramus, 2003). The results provided some evidence that infants more successfully discriminate between two language varieties when they are in different rhythm classes, but this effect did not significantly improve the model according to the Likelihood Ratio Test. Contrary to our predictions, there was no evidence in the meta-analytic results for a change with age of the importance that rhythm class plays in influencing discrimination effect sizes.

There was, however, some evidence that the effects of rhythm class change with age when it comes to language *preference*, such that infants show an increasing preference with age for a non-native language that is in the same rhythm class as the native language, but if the non-native language is in a different rhythm class, they will not show this pattern. Many discrimination tasks rely on measures that are influenced by infants' preferences (e.g., looking times), so it is useful to be aware of how these preferences change with age and may confound results. We caution that the sample is predominantly composed of experiments that tested stress-timed Germanic languages, syllable-timed Romance languages and mora-timed Japanese, so varieties grouped by rhythm class generally share more properties than just rhythm. Future studies contrasting languages from the same rhythm class, but distinct language families, will help clarifying the basis of this effect. Unlike in discrimination tasks, no durational metrics replaced the factor rhythm class in accounting for the interaction with age in preference tasks. It is possible that *regardless of rhythm*, infants show a novelty preference when the non-native variety is different only in subtle ways (e.g., intonational or segmental differences). Further evidence is needed to establish exactly which characteristics drive infants' preferences.

4.1.1. The role of rhythm class in discrimination at different ages

Both the *rhythmic class acquisition hypothesis* and the *native language acquisition hypothesis* make predictions that the effect of rhythm class will change over time. That is, newborns can discriminate between languages in different rhythm classes better than in the same rhythm class, and as age increases, infants should instead be able to discriminate any language variety from their native language. Meta-analytic results estimated that newborns' discrimination of languages in different rhythm classes will be successful. However, the claim that newborns cannot easily discriminate between two languages in the same rhythm class has been reported only by Nazzi et al. (1998), who did not find evidence that French-learning infants could discriminate between English and Dutch. Note that Ramus (2002) mentions unpublished data showing that newborns do not discriminate between syllable-timed Catalan and Spanish, but we were unable to obtain any other details about this study and so could not include it in our analyses. Our systematic review has revealed that a key means of falsifying the *rhythmic class acquisition hypothesis* – by showing newborns' ability to discriminate between languages within a rhythm class – has seldom been reported in published studies. Further studies that test newborns' discrimination of rhythmically similar languages are needed, to establish whether the body of evidence points towards a strong effect of rhythm class in newborns, and whether this effect is stronger in newborns than in later ages.

Reviews on language discrimination generally state that the ability to discriminate between two rhythmically similar languages emerges at four months (Gervain & Mehler, 2010; Höhle, Bijeljac-Babic, & Nazzi,

2020). As there was no interaction of rhythm class and age for discrimination, the meta-analytic results do not support this claim, and a closer look at individual studies reveals that further research is needed to establish when this ability emerges. Firstly, note that no included study tested infants between the ages of one week and two months. Some studies have found that infants can discriminate between rhythmically similar languages from as early as at two months (Christophe & Morton, 1998; Christophe, Nespor, Guasti, & van Ooijen, 2003). This suggests that sensitivity to within-rhythm-class differences emerge earlier than generally claimed.

Although studies testing infants between three to six months have generally found that infants can distinguish their native language from a language in the same rhythm class (Bosch & Sebastián-Gallés, 1997; Molnar et al., 2013; Molnar & Carreiras, 2015; Nácar Garcia et al., 2018; Nazzi et al., 2000; Shafer, Shucard, & Jaeger, 1999), some unpublished studies have yielded null results (Bosch, 2010; de Ruiter, Geambasu, & Levelt, 2015; Johnson, Jusczyk, & Ramus, 2003; Vicenik, 2011). Likewise, Peña et al. (2010) found no differential electrophysiological processing of native Chilean Spanish and non-native Italian in three-month-olds. Nor did Chong et al. (2018) find significant evidence that American infants could discriminate between English and German at five months, with evidence for discrimination emerging only at seven months. This qualitative consideration of results reflects the meta-analytic finding of a weak effect of rhythm class, and no interaction with age. Furthermore, relatively few studies have examined discrimination beyond six months, and none of these tested languages in different rhythm classes, further indicating there is not yet a full picture of how language discrimination skills change with age over the first year of life. Additional factors may be at play in accounting for discrimination skills, and rhythm class alone appears insufficient in explaining why infants can discriminate between certain languages at two months (Christophe et al., 2003) but not others at five months (Chong et al., 2018).

An exploratory analysis revealed a significant effect of language background, whereby bilinguals showed a smaller novelty effect in discrimination tasks, a smaller familiarity effect in preference tasks and a stronger effect of differences in rPVI-C facilitating discrimination (see 3.5). The small number of records for bilingual infants warrants caution in interpreting this effect, and in general more studies are needed that test bilingual infants. Like monolinguals, as yet, it is unclear at what age bilinguals acquiring two rhythmically similar languages learn to discriminate between them (it appears to be before 3.5 months; Molnar et al., 2013), and how much earlier this would be achieved than in monolingual infants learning one of the languages in question. Future studies that compare monolinguals and bilinguals' discrimination skills in the early months of life would be able to illuminate effects of the rhythmic properties of the tested languages, language nativeness and multilingualism.

4.1.2. The role of rhythm class in discrimination of non-native languages

The distinction between the two aforementioned hypotheses discussed by Nazzi and Ramus (2003) is that the *rhythmic class acquisition hypothesis* predicts that by a few months of age, infants should be able to discriminate between two non-native languages in their native rhythm class, while the *native language acquisition hypothesis* predicts they should not. There is currently no meta-analytic evidence one way or the other. Only two experiments have tested the right combination of languages to investigate this; two non-native languages that are in the native rhythm class; and found contrasting results. On the one hand, Nazzi et al. (2000) found that five-month-old American infants could not discriminate between German and Dutch ($g = 0.09$, supporting the *native language acquisition hypothesis*), while Johnson and Braun (2011, for which effect sizes could not be calculated) found that Canadian 4.5-month-olds could discriminate between German and Norwegian (supporting the *rhythmic class acquisition hypothesis*). This demonstrates a need for additional studies testing infants in two non-native languages that are both rhythmically similar to the infants' native language, in order to disentangle these two hypotheses.

4.1.3. The limited role of rhythm class

Overall, the meta-analytic results, taken along with a qualitative appraisal of the literature, suggest that grouping languages into rhythm classes is limited in the extent to which it accounts for language discrimination in infancy. The result that there were no effects for stimulus manipulation suggests that infants rely on rhythm as an accessible cue for language discrimination from birth. This supports earlier suggestions in the literature that infants do not need to rely on segmental cues, as they can discriminate between low-pass filtered and resynthesised language stimuli from as early as at birth (Byers-Heinlein et al., 2010; May et al., 2011; Mehler et al., 1988; Ramus, 2002; Ramus et al., 2000) and in later months (Chong et al., 2018). However, it remains unclear how crucial rhythm is in newborns' language discrimination, as only one experiment has tested newborns in rhythmically similar languages (Nazzi et al., 1998). In contrast to our predictions, based on the *rhythmic class acquisition* and *native language acquisition hypotheses* of language discrimination in infancy (Nazzi & Ramus, 2003), there was no evidence that access to rhythmic cues attenuates with age as familiarity with the native language enhances, or as sensitivity to other cues develops. The limited extent to which rhythm class accounted for variance in discrimination tasks justifies a shift of focus to durational metrics that can more sensitively account for the role of rhythm in language discrimination.

4.2. Durational cues as a predictor of language discrimination

For the second research question, we asked which durational cues best predict infants' language discrimination skills from newborns up to 12 months of age. We predicted that the effect sizes of language discrimination in younger infants would be associated with differences in consonantal and vocalic interval durational metrics. This prediction was met: larger differences in rPVI-C (the mean differences in duration between successive consonantal intervals; Grabe & Low, 2002) and smaller differences in $\triangle$V (the standard deviation of vocalic interval durations; Ramus et al., 1999) best accounted for successful language discrimination, better so than the factor of rhythm class. Again, we did not find any interaction with age that would shed light on how reliance on durational cues for discrimination changes with age.

4.2.1. Large differences in consonantal variability ease discrimination

The effect of rPVI-C was such that discrimination tended to be more successful when the tested language varieties differed in their rPVI-C measure, which suggests that infants are sensitive to successive differences in consonantal variability between language varieties they are exposed to. In comparison to $\triangle$C, rPVI-C is considered to better capture information about the stress distribution of languages (White et al., 2012) and to be less sensitive to spurious variability between speakers, sentences and registers (Grabe & Low, 2002). The significant effect of rPVI-C could reflect that the dataset involved a variety of languages that differ in their stress distributions to varying extents.

When observing individual records, some large effect sizes emerge, where the tested language varieties are in the same rhythm class but have a large difference in rPVI-C. These include Catalan and Spanish (Bosch & Sebastián-Gallés, 1997; Zacharaki & Sebastián-Gallés, 2019), Australian and South African English (Kitamura et al., 2013), Turkish and French (Christophe et al., 2003), and French and Spanish (White et al., 2016). Meanwhile, Japanese and Standard British English are in different rhythm classes but close in rPVI-C, and Ramus et al. (2000) found that newborns could not discriminate between natural recordings of these two languages. This illuminates the finding that while rhythm class has been a sufficient factor in accounting for language discrimination, identifying an acoustically-defined, continuous metric allows for a more objective and sensitive way to account for language discrimination.

In the only adult language discrimination study known to us that directly tested how differences in durational metrics accounted for successful discrimination, rPVI-C was identified as one of two of the best predictors of participants' discrimination of resynthesised and speech-rate-matched Castilian Spanish and Standard British English utterances, the other being utterance-final vocalic lengthening (nFinal-V, White et al., 2012). In a subsequent experiment of the same study testing discrimination of Standard British and Welsh Valleys English, varieties that differ less in their stress distributions, △C was found to be the best predictor. These findings suggest that sensitivity to consonantal variability does not attenuate but is maintained into adulthood and able to be recruited for language discrimination if other cues such as phonology, speech rate and the lexicon are obscured.

4.2.2. Low differences in vocalic variability ease discrimination

The effect of △V was significant but in the opposite direction to predicted, whereby discrimination effect sizes became larger as the differences in △V between the two test languages decreased. This suggests that when two tested languages differ in their vowel interval variability, it impedes sensitivity to other differences that allow successful discrimination. This finding contradicts Mehler and colleagues' prediction that large differences in vocalic intervals would facilitate infants' language discrimination, due to infants' heightened sensitivity to vowels for recognising and remembering single syllables and words in utero, at birth and over the first semester of life (Mehler et al., 1996; Nespor et al., 2003; Ramus et al., 1999).

The vowel bias may, however, be precisely the reason that differences in vowel interval durations are a poor indication of a change in language for infants. In the first study that calculated the metric △V, along with %V and △C, with the goal of establishing a purely phonetic account of language rhythm, Ramus et al. (1999) noted that vowel interval variability can be influenced by various phonological factors, including vowel reduction and contrastive vowel length. This reasoning was used to justify their finding that %V and △C were better acoustic metrics than △V for making up rhythm class (despite its name, %V expresses the ratio of both vowels *and* consonants, and so is not a purely vocalic measure). This accords with the finding that differences in △V should be low in order for discrimination to be successful; if infants begin life with a sensitivity to rhythm that can be conceived of purely in acoustic terms, then phonologically-determined differences in vowel interval variability will be heard as noise, rather than facilitating discrimination.

This is not to suggest that infants pay no attention to information carried by vowels when discriminating between languages. Pitch differences are carried in vowel intervals, and intonation has been found to play a role in language discrimination in newborns (Ramus, 2002), seven-month-olds (Chong et al., 2018) and adults (Arvaniti & Rodriquez, 2013; Hagmann & Dellwo, 2014; Vicenik & Sundara, 2013). Prominence within phonological phrases, distinguishing between heads and complements, is another feature that manifests in duration and pitch cues, and is thus carried more strongly by vowels (Langus et al., 2017). Prominence has been suggested to account for two-month-olds' ability to discriminate between head-initial-French and head-final-Turkish (Christophe et al., 2003). The present results only go so far as to suggest that when considering duration at the segmental level alone, the combination of larger differences in duration of consonant intervals and smaller differences in duration of vowel intervals seem to facilitate discrimination.

It is possible that the inverse association between differences in △V and discrimination effect sizes reflects infants' sensitivity to speech rate. Namely, infants may need a relatively consistent speech rate between languages in order to detect differences in consonantal variability, and differences in vocalic duration variability might increase as differences in speech rate increase (we thank an anonymous reviewer for pointing this out). We excluded speech rate as a durational metric from our analyses as it was available only for 59% records, however in a subset analysis of those records, we found that speech rate did not arise as a significant factor. We also note that of 107 discrimination records, 65 were from studies that reported matching their stimuli for number of syllables (a further 15 were accent discrimination studies, so differences in speech rate are expected to be minimal). Still, we do not discount the possibility that the observed effect of △V would more accurately be reflected in differences in speech rate if the data were available to us, and future research that calculates durational metrics directly from their stimuli could shed more light on this question.

4.3. Limitations and strengths

Although the reported results arose as significant, note that some were relatively small effect sizes, or had rather large confidence intervals, indicating imprecise estimates. rPVI-C and △V yielded small effect sizes, while the effect of rhythm class and age in preference tasks was small-to-medium, all with rather large 95% CIs. Moreover, the indirectness of standardised effect sizes as a measure of discrimination should be acknowledged; already infant studies must assume that an indirect measure such as looking time indicates discrimination, and standardised effect sizes are another abstraction from the core construct in question. Taken together, further and direct investigation of all findings reported here is warranted.

In compiling the dataset, certain assumptions regarding study details were made that may limit the validity of the results. Many studies do not indicate the exact language variety acquired by the infant participants or spoken by the stimuli speakers, so if this information could not be obtained, we assumed the standard variety was used. This may be problematic considering how different varieties of the same language can differ in their durational properties (Clopper & Smiljanic, 2015).

In most cases (except for Butler et al., 2011; White et al., 2014) the durational metrics were not calculated from the stimuli that were used in the discrimination stimuli. Durational metrics tend to show large amounts of between-speaker, −utterance and -register heterogeneity (Barry, Andreeva, Russo, Dimitrova, & Kostadinova, 2003; Dellwo, 2006; Loukina, Rosner, Kochanski, Keane, & Shih, 2013; Pettorino & Pellegrino, 2016; Prieto et al., 2012), so the reported metrics likely include these sources of noise. The studies that reported their own durational metrics have the advantage that we know the actual rhythmic properties of the stimuli to which the participating infants were exposed. Conversely, while precise measures of the acoustic stimuli may provide a more accurate description of the stimuli used in the experiment, they may be less useful for generalizing to the language as a whole. In any case, we used whichever metrics were available to us at the time. We urge authors of future language discrimination studies to calculate and report the durational metrics of their stimuli (including those from White et al., 2012, White et al., 2014), that we excluded from our study due to missing values), and clearly report their methods for doing so, to facilitate later evidence synthesis.

Our approach was to include all available evidence on language discrimination in infancy, applying broad inclusion criteria. Measures of consistency indicate that between-study variance is large, and that a large amount of unexplained variance remains in the data. Considering the significant effect of method in the first analysis and inspection of funnel plots, this may be partly due to the conflation of neurophysiological and behavioural methods. Neurophysiological effect sizes appear to switch between familiarity and novelty effects more readily than behavioural measures. There were no clear effects of age, stimulus manipulation or native language, which, from a qualitative overview of the literature, are likely to somewhat influence discrimination. In some cases, the participant groups were heterogeneous, which could be obscuring effects (some monolingual and some bilingual, Bahrick & Pickens, 1988; different native languages, Bosch & Sebastián-Gallés, 1997; Byers-Heinlein et al., 2010; Kinzler et al., 2007; Moon et al., 1993; Nácar Garcia et al., 2018; Zacharaki & Sebastián-Gallés, 2019). However, because all accessible evidence is compiled in the dataset, and

between the quantitative and qualitative analyses, the available body of evidence has been thoroughly synthesised. As new evidence emerges, the data could be subset, or sensitivity analyses run, to confirm that the identified sources of potential noise do not overly influence the present results. We facilitate such evidence cumulation by making our dataset publicly available.

Funnel plot asymmetry suggests that reporting bias may be present in the data. Another strength of this study, however, was the inclusion of unpublished file-drawer data and theses, and conference papers. This allows for thorough consideration of whether patterns in results of published and commonly-cited studies reflect the full body of literature or may be subject to publication and reporting biases. The collection and synthesis of durational metrics for many of the language varieties is another strength of this study. In both cases, the database is available for the addition of new data, meaning that results and interpretations can easily be updated as more infant discrimination studies are conducted, and as more durational metrics are calculated.

4.4. Future studies

Various gaps in the literature have been identified, where additional studies would address outstanding questions on the development of language discrimination in infancy. Future studies should test newborns in two rhythmically similar languages, replicating the only other known experiment that has done so (Nazzi et al., 1998), to confirm whether infants can only discriminate between rhythmically distinct languages at birth. Moreover, testing older infants' ability to discriminate between two non-native languages that are rhythmically similar to both the native language and each other (Johnson & Braun, 2011; Nazzi et al., 2000) would help to disentangle the *rhythmic class acquisition* and *native language acquisition hypotheses* (Nazzi & Ramus, 2003) in indicating whether or not infants are able to discriminate between two non-native languages in their native rhythm class.

Our finding that larger differences in rPVI-C and smaller differences in △V are better measures for accounting for discrimination than the factor rhythm class needs to be confirmed in studies that directly measure these metrics from their stimuli, and that calculate the extent to which these metrics account for infants' language discrimination (following the approach of White et al., 2014). The effect of △V emerged in the opposite direction as predicted, which warrants confirmation of the finding that discrimination is facilitated when languages are similar in the variation of their vowel interval durations. Such studies could test combinations of languages that isolate differences in just one of these metrics (e.g. varieties of a single language that differ rhythmically like syllable-timed Ghanaian English with stress-timed varieties of English; Boll-Avetisyan, Omane, & Kügler, 2020, or manipulated stimuli that differ along just one of these dimensions e.g. Paillereau et al., 2021).

One study goal was to create a Community Augmented Meta-analysis (CAMA, Cristia, Tsuji, & Bergmann, 2020; Tsuji et al., 2014) of language discrimination studies (available at http://metalab.stanford.edu/dataset/langdiscrim/). Any future studies in infant language discrimination are invited to contribute data to the CAMA, as well as to conduct further meta-analyses. For instance, exploratory analysis did not show a clear effect of language nativeness, but anecdotally, it seems that language nativeness does play some kind of role in discrimination (Butler et al., 2011; Christophe & Morton, 1998; Chung, 2002; Kitamura, Panneton, Diehl, & Notley, 2006; Kitamura et al., 2013; Mehler et al., 1988; Nazzi et al., 2000). More sophisticated analyses could take into consideration variables such as the amount of familiarity to a non-native variety, acoustic distance of tested languages from the native variety along segmental and suprasegmental dimensions, and language dominance in the case of multilinguals, to further establish the role than language nativeness plays over the first year of life.

Certain understudied languages needed to be excluded from the durational metric analyses because metrics could not be obtained (Basque, Tagalog, Quebecois French, Cuban Spanish, New York, and Miami Hispanic English). Various durational metrics were also excluded due to missing data (utterance-final lengthening, speech rate and other measures of interval durations and variability included in White et al., 2012, White et al., 2014). Therefore, we could not provide evidence on the role of speech rate or utterance-final lengthening, which were significant factors in White et al. (2014) and White et al. (2012). Future discrimination studies should calculate and report these metrics on their stimuli to better establish the amount of variance explained by these metrics and observe whether the present findings are maintained when new language varieties are included in the analyses. The role of intonation is another promising factor not accounted for in this study. Vicenik and Sundara (2013) quantified certain aspects of intonation which could be calculated and reported for future investigation of the role of intonation in language discrimination.

5. Conclusions

In this systematic review of infants' language discrimination skills from birth to 12 months of age, meta-analytic evidence shows that infants more easily discriminate between language varieties that differ in their successive consonant interval variability but do not differ in their global vowel interval variability. This study supports previous research in highlighting the importance of rhythm in language discrimination, but brings the novel finding that rhythm can be operationalised as the mean difference in duration of successive consonant intervals (rPVI-C, Grabe & Low, 2002) and the standard deviation of vowel interval durations (△V, Ramus et al., 1999), which account for variance in discrimination more reliably than does rhythm class. The present results do not support previous hypotheses on language discrimination regarding changes with age; no developmental trajectory was identified whereby sensitivity to, or reliance on, durational cues attenuates with age. An adult study showed similar results to the present findings, suggesting that consonant interval variability, operationalised as rPVI-C, remains a useful cue in adulthood (White et al., 2012). Nor do the present findings support previous predictions that large differences in vowel interval variability are important for language discrimination in newborns (Mehler et al., 1996; Nespor et al., 2003; Ramus et al., 1999). Instead, the results indicate that over the course of infancy, small differences in vowel interval variability and large differences in consonantal interval variability are the optimal durational cues that facilitate language discrimination.

Registry

This meta-analysis was first registered with Open Science Framework (OSF) on 16/04/2020 (available at https://osf.io/396yb).

Funding

This work was supported by the Erasmus Mundus Joint Master Degree scholarship provided by the European Commission and the University of Tokyo Excellent Young Researcher Startup Fund.

Declaration of Competing Interest

None.

Acknowledgements

Thank you to those who provided data and study details for the meta-analysis: Laura Bosch, Krista Byers-Heinlein, Adam Chong, Anne Christophe, Alejandrina Cristia, Ghislaine Dehaene-Lambertz, Christine Kitamura, Kristien de Ruiter, Andreea Geambasu, Carlos Guerrero-Mosquera, Hilary Killam, Claartje Levelt, Yasuyo Minagawa, Monika Molnar, Loreto Nácar Garcia, Thierry Nazzi, Melissa Paquette-Smith,

Hiroki Sato, Valerie Shafer, Melanie Soderstrom, Megha Sundara, Janet Werker and Konstantina Zacharaki and to all other scientists who contacted us following our call for studies. For their insightful discussion thank you to Christina Bergmann, Audrey Bürki, Chiara Cantiani, Kateřina Chládková, Vânia de Aguiar, Zoë Firth, Claudia Männel, Fódhla Ní Chéileachair, Srdjan Popov, Hugh Rabagliati, Jessica Ramos Sanchez, Sien van der Plank, two anonymous reviewers, and attendees of the 7th Summer Neurolinguistics School, Moscow, and Many Paths to Languages, 2020.

Appendix A. PRISMA Checklist

Table A.1
PRISMA Checklist. From Moher et al. (2009). For more information, visit: www.prisma-statement.org.

Section/topic	#	Checklist item	Section no.
Title			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	Title page
Abstract			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	Abstract
Introduction			
Rationale	3	Describe the rationale for the review in the context of what is already known.	1.3
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	1.4
Method			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	2
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	2.1
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	2.2
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	2.2
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	2.2
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	2.3, 2.4
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	2.3
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	–
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	2.4
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I^2) for each meta-analysis.	2.5
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	3.6
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	2.5.2–2.5.5
Results			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	2.2
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	Appendix C
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	–
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	Appendix D, Appendix E
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	3.1
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	3.6
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	3.2–3.5
Discussion			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	4
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	4.3
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	5
Funding			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	Funding

Appendix B. Durational metrics

Table B.1
Durational metrics of all language varieties and original sources. UC = unclassified.

	Rhythm class	%V	△C	△V	VarcoV	nPVI-V	rPVI-C	References
Basque	syllable	47.58	39.18	36.60	44.28			Molnar, Carreiras, & Gervain, 2016; Nespor et al., 2011
Catalan	syllable	48.42	35.98	33.66	38.22	37.35	63.69	Gavaldà-Ferré, 2007; Grabe & Low, 2002; Prieto et al., 2012; Ramus et al., 1999
Catalan, Eastern	syllable	48.48	34.16	30.02	38.22	37.35	67.8	Grabe & Low, 2002; Prieto et al., 2012; Ramus et al., 1999
Catalan, Western	syllable	48.19	42.84	47.41			63.27	Gavaldà-Ferré, 2007
Dutch	stress	41.67	50.52	46.50	64.45	77.65	52.80	Grabe & Low, 2002; Mairano, 2011; Pettorino & Pellegrino, 2016; Ramus et al., 1999; White & Mattys, 2007
English	stress	51.68	55.29	56.34	56.79	61.32	62.96	Arvaniti, 2012; Boll-Avetisyan et al., 2020; Clopper & Smiljanic, 2015; Coetzee & Wissing, 2007; Ding & Xu, 2016; Grabe & Low, 2002; Kawase, Kim, & Davis, 2016; Lee, Kitamura, Burnham, & Todd, 2014; Loukina et al., 2013; Mukai & Tucker, 2015; Pettorino & Pellegrino, 2016; Prieto et al., 2012; Ramus et al., 1999; Romano, 2010; Vicenik & Sundara, 2013; White et al., 2012; White et al., 2014; White & Mattys, 2007
English, Trans-Atlantic	stress	51.72	55.24	56.43	56.91	60.90	63.22	Arvaniti, 2012; Boll-Avetisyan et al., 2020; Clopper & Smiljanic, 2015; Ding & Xu, 2016; Grabe & Low, 2002; Loukina et al., 2013; Mukai & Tucker, 2015; Pettorino & Pellegrino, 2016; Prieto et al., 2012; Ramus et al., 1999; Vicenik & Sundara, 2013; White et al., 2012; White et al., 2014; White & Mattys, 2007
English, American	stress	49.61	57.02	56.82	55.25	58.54	67.53	Arvaniti, 2012; Boll-Avetisyan et al., 2020; Clopper & Smiljanic, 2015; Mukai & Tucker, 2015; Ramus et al., 1999; Vicenik & Sundara, 2013
English, American, Canadian	stress	49.12	56.81	53.31	53.46	57.01	67.25	Clopper & Smiljanic, 2015; Grenon & White, 2008; Mukai & Tucker, 2015
English, American, New England	stress	49.74	56.79	55.55	55.70	58.15	67.61	Boll-Avetisyan et al., 2020; Clopper & Smiljanic, 2015
English, American, Mid-Atlantic	stress	50.25	56.85	57.77	56.18	58.95	66.86	Clopper & Smiljanic, 2015
English, American, Midland	stress	50.27	59.46	59.64	55.7	58.72	70.4	Clopper & Smiljanic, 2015
English, American, North	stress	49.4	56.66	52.97	53.38	57.4	67.08	Clopper & Smiljanic, 2015
English, American, West	stress	48.16	57.90	53.65	54.03	57.62	67.68	Arvaniti, 2012; Clopper & Smiljanic, 2015; Vicenik & Sundara, 2013
English, American, Miami Hispanic	stress	49.32	55.2					Enzinna, 2016
English, American, New York Hispanic	stress					43.57		Shousterman, 2014
English, Australian	stress	43.71	64.7	46.39	51.72	71.15	49.4	Kawase et al., 2016; Lee et al., 2014; Romano, 2010; Vicenik & Sundara, 2013
English, British	stress	53.23	63.82	55.25	58.06	68.43	60.09	Ding & Xu, 2016; Grabe & Low, 2002; Loukina et al., 2013; Pettorino & Pellegrino, 2016; Prieto et al., 2012; White et al., 2012; White & Mattys, 2007
English, British, Standard	stress	54.83	40.62	39.56	58.51	57.79	55.97	Grabe & Low, 2002; Loukina et al., 2013; Pettorino & Pellegrino, 2016; Prieto et al., 2012; White et al., 2012; White & Mattys, 2007
English, British, West Country	stress	45.9	73.5	63.9	56.5	72.5	81.5	White et al., 2014
English, British, Welsh	stress	44.5	72.7	60.6	52.3	74.6	79.1	White et al., 2014
English, British, Scottish	stress	44.5	78.3	65.1	56	71.7	91.3	White et al., 2014
English, South African	stress					60.66	68.15	Coetzee & Wissing, 2007
English, Chinese-accented	UC	47.13	63.5	53.4	47.2	48.81	71.62	Ding & Xu, 2016
English, French-accented	UC	45	74.5	63.1	56.6	68.4	84.4	White et al., 2014
English, Spanish-accented	UC	41	57	47	54	66	65	White & Mattys, 2007
Finnish	syllable	48.45	45.90	50.22	50.58	47.80	51.64	Mairano, 2011; Mairano & Romano, 2011, Mairano & Romano, 2011b; Nespor et al., 2011; Romano, 2010
French, Standard	syllable	51.06	45.35	40.97	67.74	49.07	60.63	Dellwo, 2006; Grabe & Low, 2002; Loukina et al., 2013; Pettorino & Pellegrino, 2016; Ramus et al., 1999; White & Mattys, 2007
French, Quebecois	syllable	51.78	47.57	41.41	50.42	48.41		Cichocki, Selouani, & Perreault, 2014; Kaminskaïa, 2020; Lidji, Palmer, Peretz, & Morningstar, 2011; Romano, 2010; Roy, Macoir, Martel-Sauvageau, & Boudreault, 2012
French, English-accented	UC			62.98				Vieru, de Mareüil, & Adda-Decker, 2011
German	stress	40.40	63.38	45.62	51.5	53.90	66.50	Arvaniti, 2012; Barry et al., 2003; Dellwo, 2006; Grabe & Low, 2002; Mairano & Romano, 2011; Pettorino & Pellegrino, 2016; Vicenik & Sundara, 2013
Italian	syllable	48.82	46.03	41.75	53.30	47.54	52.07	Arvaniti, 2012; Barry et al., 2003; Mairano & Romano, 2011; Pettorino & Pellegrino, 2016; Ramus et al., 1999; White, Payne, & Mattys, 2009
Japanese	mora	52.66	41.51	41.03	56	38.73	50.49	Grabe & Low, 2002; Grenon & White, 2008; Mairano & Romano, 2011; Pettorino & Pellegrino, 2016; Ramus et al., 1999
Mandarin, Mainland	syllable	55.46	43.69	46.50	45.56	48.76	49.82	Ding & Xu, 2016; Grabe & Low, 2002; Lin & Wang, 2007; Mairano, 2011; Mok, 2009
Mandarin, Taiwanese	syllable	68.21			59.88		61.3	Loukina et al., 2013
Russian	stress	55.25	52.74	35.27	65.31	44.99	53.14	Loukina et al., 2013; Mairano, 2011; Mairano & Romano, 2011; Pettorino & Pellegrino, 2016; Stojanovic, 2013
Spanish	syllable	47.45	38.63	26.03	44.41	42.25	52.15	

(*continued on next page*)

Table B.1 (*continued*)

	Rhythm class	%V	△C	△V	VarcoV	nPVI-V	rPVI-C	References
								Arvaniti, 2012; Grabe & Low, 2002; Pettorino & Pellegrino, 2016; Prieto et al., 2012; Ramus et al., 1999; Toledo, 2009; White et al., 2012; White & Mattys, 2007
Spanish, Castilian	syllable	46.84	31.04	26.03	36.94	36.92	45.1	Grabe & Low, 2002; Pettorino & Pellegrino, 2016; Prieto et al., 2012; Ramus et al., 1999; White et al., 2012; White & Mattys, 2007
Spanish, Latin American	syllable	48.25	46.97		53.3	47.55	53.7	Arvaniti, 2012; Toledo, 2009
Spanish, Latin American, Cuban	syllable	42.71	48.61			40.68		Toledo, 2009
Tagalog	syllable	46.8	46.6	45.2				Bird et al., 2005
Turkish	syllable	46.62	54.00	38.78	43.15	47.01	67.22	Mairano, 2011; Mairano & Romano, 2011; Nespor et al., 2011; Romano, 2010; Stojanovic, 2013
Catalan, Eastern/ Spanish, Castilian	syllable	47.66	32.60	28.02	37.58	37.13	56.45	Grabe & Low, 2002; Pettorino & Pellegrino, 2016; Prieto et al., 2012; Ramus et al., 1999; White et al., 2012; White & Mattys, 2007
English, American/ French, Standard	UC	50.33	51.19	48.89	61.49	53.80	64.08	Arvaniti, 2012; Boll-Avetisyan et al., 2020; Clopper & Smiljanic, 2015; Dellwo, 2006; Grabe & Low, 2002; Loukina et al., 2013; Mukai & Tucker, 2015; Pettorino & Pellegrino, 2016; Ramus et al., 1999; Vicenik & Sundara, 2013; White & Mattys, 2007
English, French-accented/French, English-accented	UC			63.04				Vieru et al., 2011; White et al., 2014

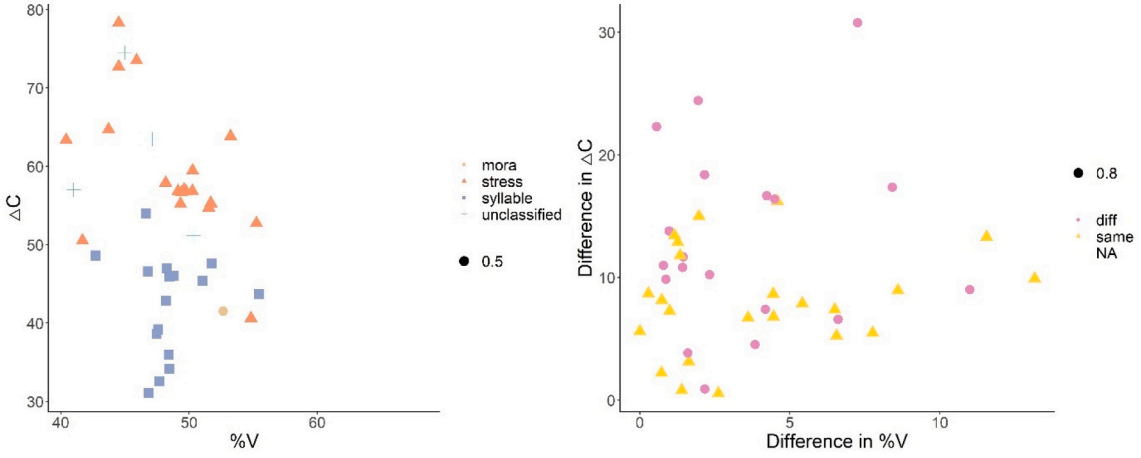


Fig. B.1. Plots of %V by △C.
See Grabe and Low (2002), Nespor et al. (2011), Ramus et al. (1999).

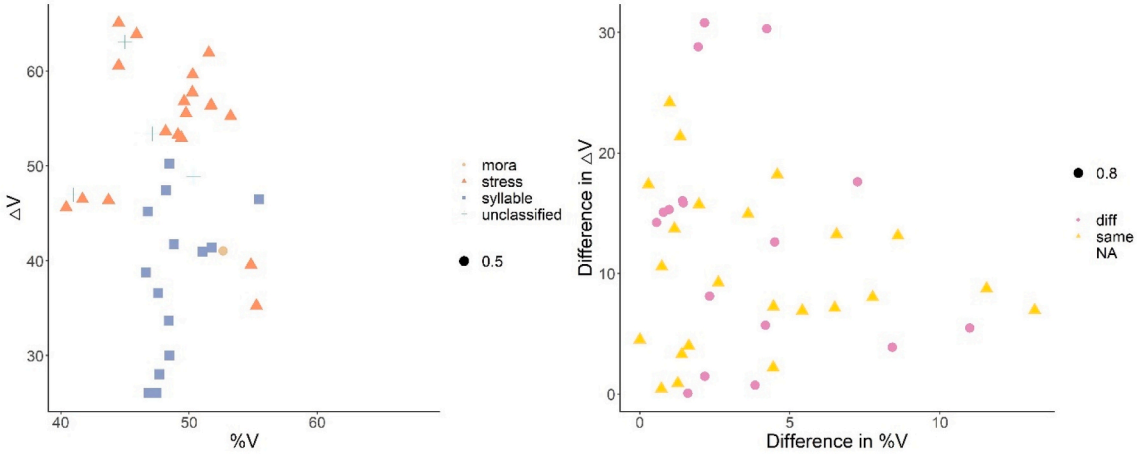


Fig. B.2. Plots of %V by △V.
See Ramus et al. (1999).

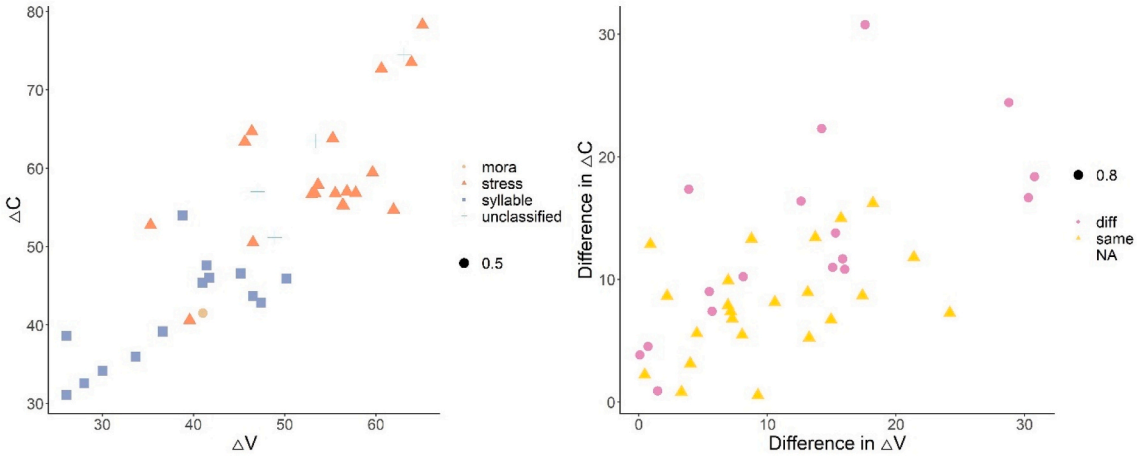


Fig. B.3. Plots of △V by △C.
See Ramus et al. (1999).

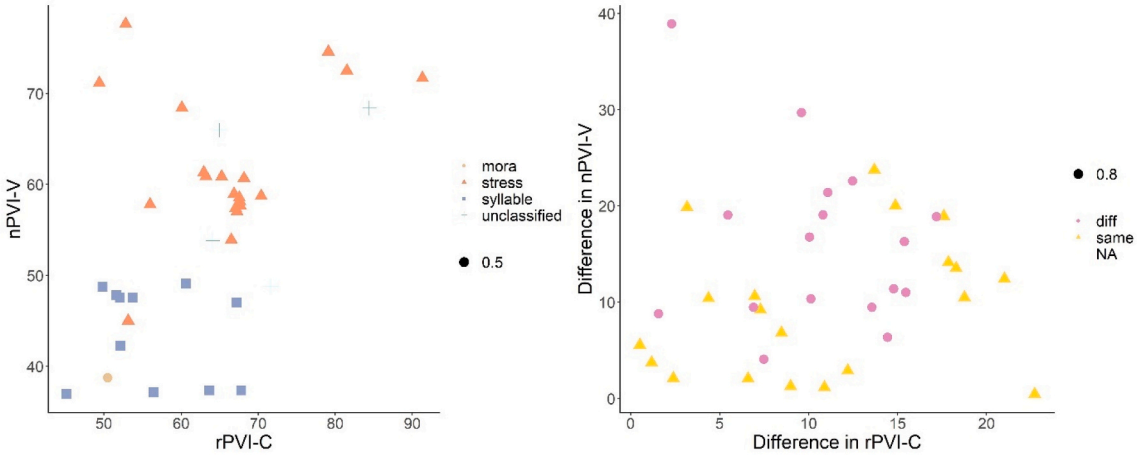


Fig. B.4. Plots of rPVI-C by nPVI-V.
See Grabe and Low (2002), White and Mattys (2007).

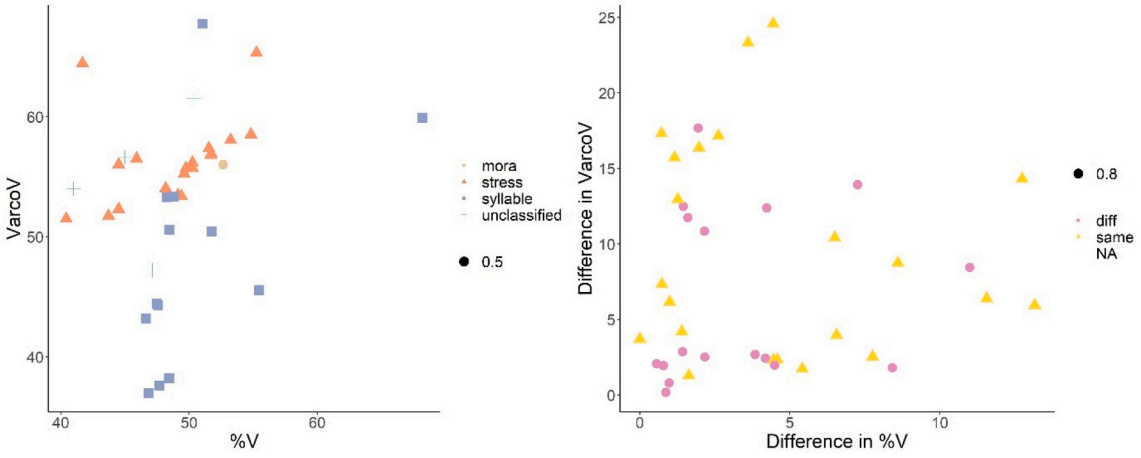


Fig. B.5. Plots of %V by VarcoV.
See (White & Mattys, 2007).

Appendix C. Study characteristics

Table C.1
Study characteristics.

Authors (year)	Task	Age (months)	Native language(s)	Tested languages	Rhythm class difference	Stimulus manipulation	Method
Bahrick and Pickens (1988)*	discrimination	5	English, Spanish	English, Spanish	Different	None	CF
Bosch (2010)**	preference, discrimination	4, 6, 9	Eastern Catalan, Spanish	Eastern Catalan, Western Catalan, Spanish, Basque	Same	None	CF
Bosch and Sebastián-Gallés (1997)	preference	4	Catalan, Spanish	Catalan, Spanish, English, Italian	Same, different	None, LPF	HPP
Bosch, Cortés, and Sebastián-Gallés (2001)	preference	4	Catalan, Spanish	Catalan, Spanish	Same	None	HPP
Bosch and Sebastián-Gallés (2001)	discrimination	4	Catalan, Spanish	Catalan, Spanish	Same	None	HPP
Butler et al. (2011)	discrimination	5, 7	West Country English	West Country English, Scottish English, Welsh English	Same	None	HPP
Byers-Heinlein et al. (2010)*	preference, discrimination	0	English, Tagalog, Chinese	English, Tagalog	Different	LPF	HAS
Chong et al. (2018)	discrimination	5, 7	English	English, German	Different	None, LPF, monotone, f0-matched	HPP
Christophe et al. (2003)	discrimination	2	French	French, Turkish	Same	Resynthesised	HAS
Christophe and Morton (1998)	discrimination	2	English	English, French, Dutch, Japanese	Same, different	None	HAS
Chung (2002)*	discrimination	4, 10	Pittsburgh English	Pittsburgh English, New York Hispanic English, Chinese-accented English, Mainland Mandarin, Taiwanese Mandarin	Same, UC	None	CF
Cristia et al. (2014)*	discrimination	3, 5	Parisian French	Parisian French, Quebecois French	Same	None	NIRS
Dehaene-Lambertz and Houston (1998)	preference	2	English, French	English, French	Different	None, LPF	HPP
de Ruiter et al. (2015)	preference	7, 8	Dutch	Dutch, English	Same	None	HPP
Diehl et al. (2006)**	preference	6, 8	American English	American English, Australian English	Same	None	HR
Fava, Hull, and Bortfeld (2014)	preference	5, 8, 12	English	English, Spanish	Different	None	NIRS
Hayashi, Tamekawa, and Kiritani (2001)	preference	5, 8, 10	Japanese	Japanese, English	Different	None	HPP
Johnson et al. (2003)	discrimination	5	English	English, Dutch	Same	Resynthesised, LPF	HPP
Johnson and Braun (2011)**	discrimination	4	English	English, German, Norwegian	Same	None	HPP
Kinzler et al. (2007)	preference	6, 10	English, Spanish, French	English, Spanish, French, English-accented French, French-accented English	Different, UC	None	CF, FC
Kitamura et al. (2006)	preference	6	American English	American English, Australian English	Same	None	CF
Kitamura et al. (2013)	preference, discrimination	3, 6, 9, 10	Australian English	Australian English, American English, South African English	Same	None	CF
May et al. (2011)*	preference	0	English	English, Tagalog	Different	LPF, backwards	NIRS
Mehler et al. (1988)	discrimination	0, 2	French, English	French, Russian, English, Italian	Different	None, LPF, backwards	HAS, CF
Minagawa-Kawai et al. (2011)	preference	4	Japanese	Japanese, English	Different	None	NIRS
Molnar and Carreiras (2015)*	discrimination	3	Basque, Spanish	Basque, Spanish	Same	None	CF
Molnar et al. (2013)*	discrimination	3	Basque, Spanish	Basque, Spanish	Same	LPF	CF
Moon et al. (1993)	preference	0	English, Spanish	English, Spanish	Different	None	HAS
Nácar Garcia et al. (2018)	discrimination	4	Catalan, Spanish	Catalan, Spanish, Italian, German	Same, different	None	EEG
Nazzi et al. (2000)	discrimination	5	English	English, Italian, Japanese, Dutch, German, Spanish	Same, different	None	HPP
Nazzi et al. (1998)	discrimination	0	French	English, Dutch, Japanese	Same, different	None	HAS
Paquette-Smith and Johnson (2015)	discrimination	5	English	English, Spanish-accented English, Spanish	Different, UC	None	HPP
Peña et al. (2010)**	discrimination	3, 9	Spanish	Spanish, Italian, Japanese	Same, different	None	EEG
Ramus et al. (2000)	discrimination	0	French	Dutch, Japanese	Different	None, resynthesised, backwards	HAS
Ramus (2002)	discrimination	0	French	Dutch, Japanese	Different		HAS

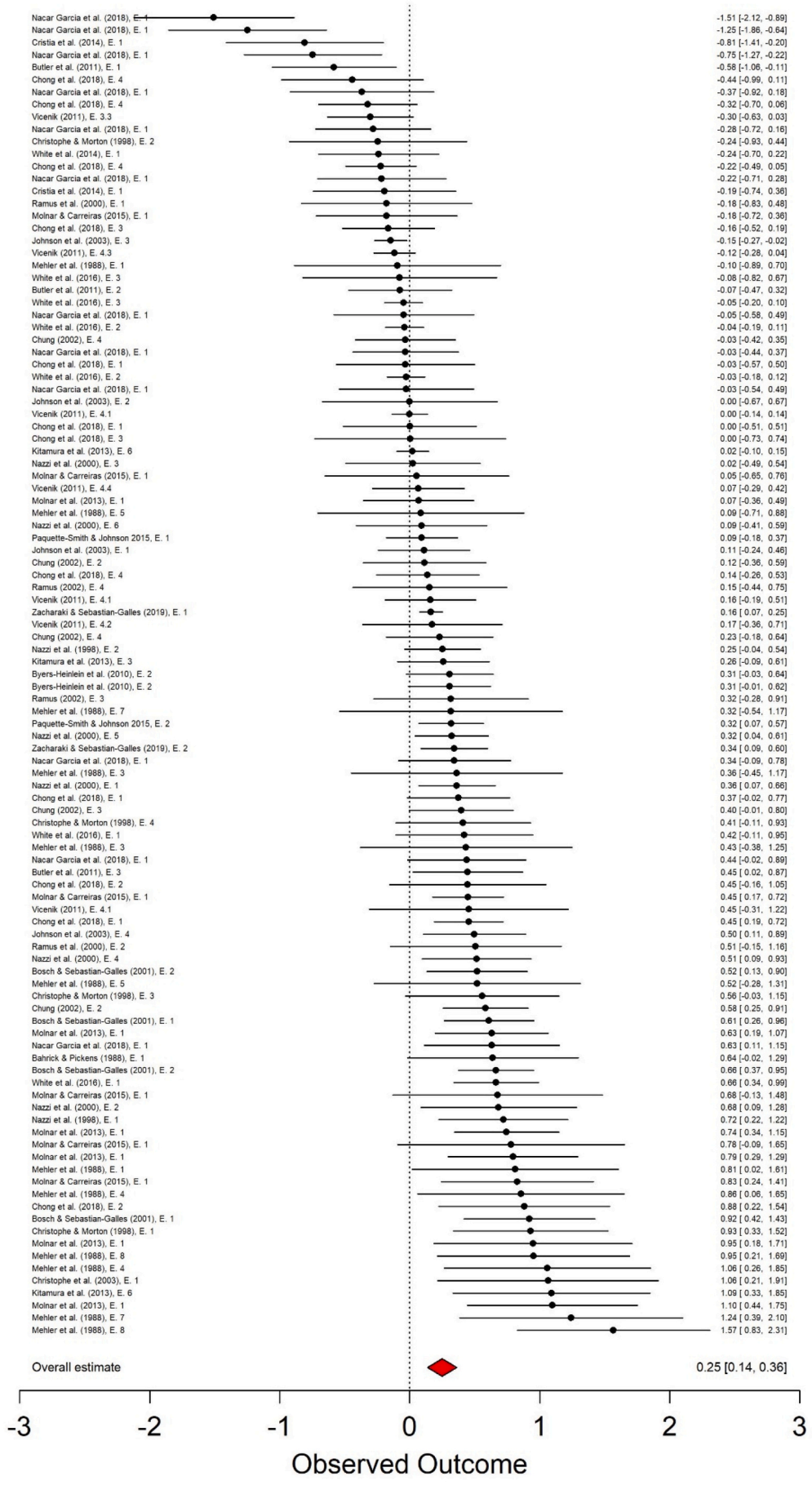
(*continued on next page*)

Table C.1 (*continued*)

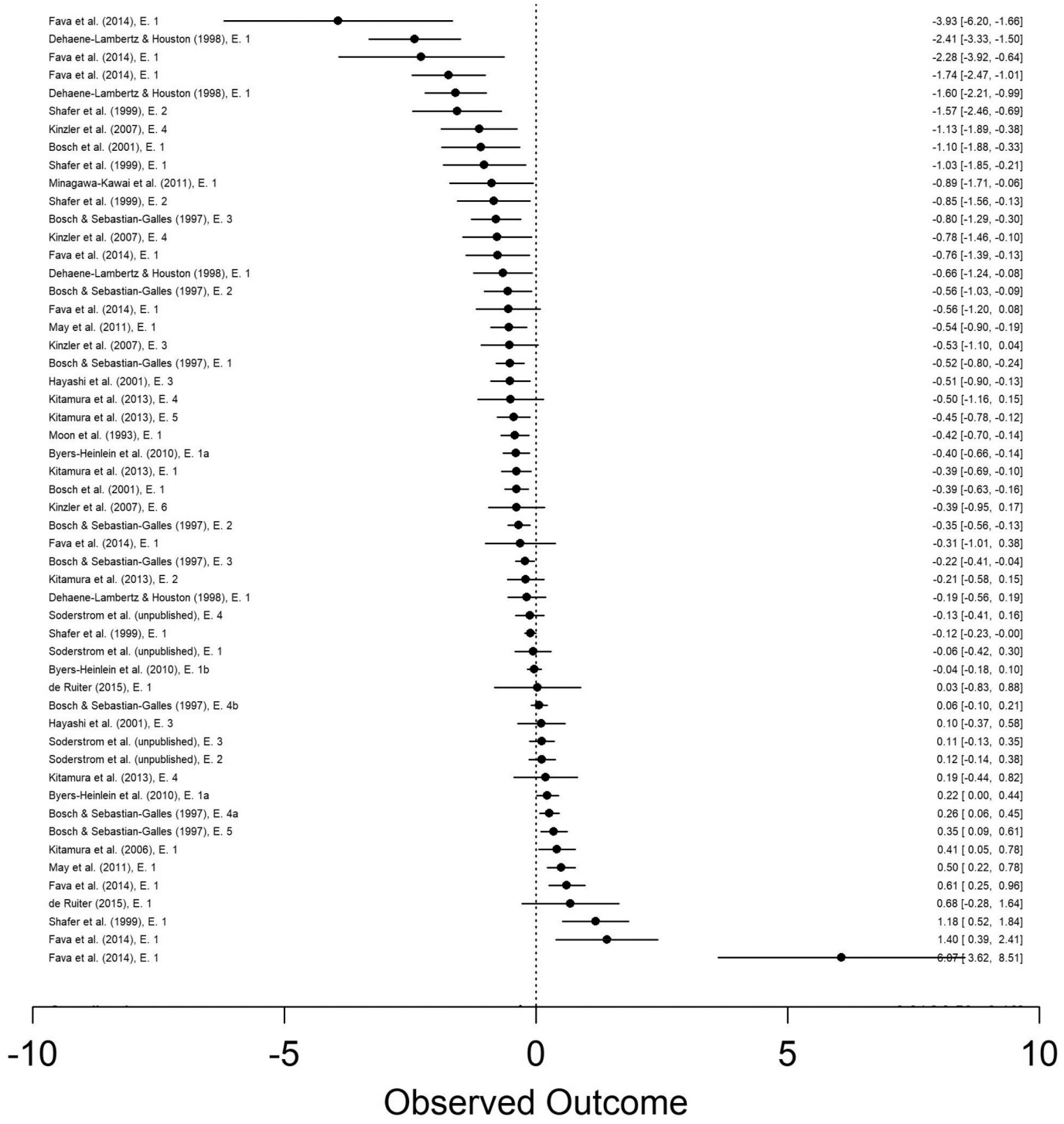
Authors (year)	Task	Age (months)	Native language(s)	Tested languages	Rhythm class difference	Stimulus manipulation	Method
						Resynthesised, intonation-matched	
Sato et al. (2012)**	preference	0	Japanese	Japanese, English	Different	None	NIRS
Shafer et al. (1999)	preference	3	English	English, Italian, Dutch	Same, different	None	EEG
Soderstrom, Xu Rattanasone, Demuth, and Seidl (2021)	preference	5, 7	Canadian English	Canadian English, Australian English	Same	None, LPF	HPP
Vicenik (2011)	discrimination	5, 7, 9	American English	German, Australian English, Dutch, Japanese	Same, different	None, intonation-matched	HPP
White et al. (2014)	discrimination	7	West Country English	West Country English, French-accented English	Uc	None	HPP
White et al. (2016)	discrimination	5	English	Finnish, French, Spanish	Same	None	HPP
Zacharaki and Sebastián-Gallés (2019)*	discrimination	4	Catalan, Spanish	Eastern Catalan, Western Catalan, Spanish	Same	None	HPP

**not included in meta-analysis, *not included in durational metrics analysis, CF = central fixation, EEG = electroencephalography, FC = forced choice, HAS = high amplitude sucking, HPP = head-turn preference paradigm, LPF = low-pass filtered, NIRS = near-infrared spectroscopy, UC = unclassified.

Appendix D. Forest plot for discrimination tasks


Nacar Garcia et al. (2018), E. 1 -1.51 [-2.12, -0.89]
Nacar Garcia et al. (2018), E. 1 -1.25 [-1.86, -0.64]
Cristia et al. (2014), E. 1 -0.81 [-1.41, -0.20]
Nacar Garcia et al. (2018), E. 1 -0.75 [-1.27, -0.22]
Butler et al. (2011), E. 1 -0.58 [-1.06, -0.11]
Chong et al. (2018), E. 4 -0.44 [-0.99, 0.11]
Nacar Garcia et al. (2018), E. 1 -0.37 [-0.92, 0.18]
Chong et al. (2018), E. 4 -0.32 [-0.70, 0.06]
Vicenik (2011), E. 3.3 -0.30 [-0.63, 0.03]
Nacar Garcia et al. (2018), E. 1 -0.28 [-0.72, 0.16]
Christophe & Morton (1998), E. 2 -0.24 [-0.93, 0.44]
White et al. (2014), E. 1 -0.24 [-0.70, 0.22]
Chong et al. (2018), E. 4 -0.22 [-0.49, 0.05]
Nacar Garcia et al. (2018), E. 1 -0.22 [-0.71, 0.28]
Cristia et al. (2014), E. 1 -0.19 [-0.74, 0.36]
Ramus et al. (2000), E. 1 -0.18 [-0.83, 0.48]
Molnar & Carreiras (2015), E. 1 -0.18 [-0.72, 0.36]
Chong et al. (2018), E. 3 -0.16 [-0.52, 0.19]
Johnson et al. (2003), E. 3 -0.15 [-0.27, -0.02]
Vicenik (2011), E. 4.3 -0.12 [-0.28, 0.04]
Mehler et al. (1988), E. 1 -0.10 [-0.89, 0.70]
White et al. (2016), E. 3 -0.08 [-0.82, 0.67]
Butler et al. (2011), E. 2 -0.07 [-0.47, 0.32]
White et al. (2016), E. 3 -0.05 [-0.20, 0.10]
Nacar Garcia et al. (2018), E. 1 -0.05 [-0.58, 0.49]
White et al. (2016), E. 2 -0.04 [-0.19, 0.11]
Chung (2002), E. 4 -0.03 [-0.42, 0.35]
Nacar Garcia et al. (2018), E. 1 -0.03 [-0.44, 0.37]
Chong et al. (2018), E. 1 -0.03 [-0.57, 0.50]
White et al. (2016), E. 2 -0.03 [-0.18, 0.12]
Nacar Garcia et al. (2018), E. 1 -0.03 [-0.54, 0.49]
Johnson et al. (2003), E. 2 0.00 [-0.67, 0.67]
Vicenik (2011), E. 4.1 0.00 [-0.14, 0.14]
Chong et al. (2018), E. 1 0.00 [-0.51, 0.51]
Chong et al. (2018), E. 3 0.00 [-0.73, 0.74]
Kitamura et al. (2013), E. 6 0.02 [-0.10, 0.15]
Nazzi et al. (2000), E. 3 0.02 [-0.49, 0.54]
Molnar & Carreiras (2015), E. 1 0.05 [-0.65, 0.76]
Vicenik (2011), E. 4.4 0.07 [-0.29, 0.42]
Molnar et al. (2013), E. 1 0.07 [-0.36, 0.49]
Mehler et al. (1988), E. 5 0.09 [-0.71, 0.88]
Nazzi et al. (2000), E. 6 0.09 [-0.41, 0.59]
Paquette-Smith & Johnson 2015, E. 1 0.09 [-0.18, 0.37]
Johnson et al. (2003), E. 1 0.11 [-0.24, 0.46]
Chung (2002), E. 2 0.12 [-0.36, 0.59]
Chong et al. (2018), E. 4 0.14 [-0.26, 0.53]
Ramus (2002), E. 4 0.15 [-0.44, 0.75]
Vicenik (2011), E. 4.1 0.16 [-0.19, 0.51]
Zacharaki & Sebastian-Galles (2019), E. 1 0.16 [0.07, 0.25]
Vicenik (2011), E. 4.2 0.17 [-0.36, 0.71]
Chung (2002), E. 4 0.23 [-0.18, 0.64]
Nazzi et al. (1998), E. 2 0.25 [-0.04, 0.54]
Kitamura et al. (2013), E. 3 0.26 [-0.09, 0.61]
Byers-Heinlein et al. (2010), E. 2 0.31 [-0.03, 0.64]
Byers-Heinlein et al. (2010), E. 2 0.31 [-0.01, 0.62]
Ramus (2002), E. 3 0.32 [-0.28, 0.91]
Mehler et al. (1988), E. 7 0.32 [-0.54, 1.17]
Paquette-Smith & Johnson 2015, E. 2 0.32 [0.07, 0.57]
Nazzi et al. (2000), E. 5 0.32 [0.04, 0.61]
Zacharaki & Sebastian-Galles (2019), E. 2 0.34 [0.09, 0.60]
Nacar Garcia et al. (2018), E. 1 0.34 [-0.09, 0.78]
Mehler et al. (1988), E. 3 0.36 [-0.45, 1.17]
Nazzi et al. (2000), E. 1 0.36 [0.07, 0.66]
Chong et al. (2018), E. 1 0.37 [-0.02, 0.77]
Chung (2002), E. 3 0.40 [-0.01, 0.80]
Christophe & Morton (1998), E. 4 0.41 [-0.11, 0.93]
White et al. (2016), E. 1 0.42 [-0.11, 0.95]
Mehler et al. (1988), E. 3 0.43 [-0.38, 1.25]
Nacar Garcia et al. (2018), E. 1 0.44 [-0.02, 0.89]
Butler et al. (2011), E. 3 0.45 [0.02, 0.87]
Chong et al. (2018), E. 2 0.45 [-0.16, 1.05]
Molnar & Carreiras (2015), E. 1 0.45 [0.17, 0.72]
Vicenik (2011), E. 4.1 0.45 [-0.31, 1.22]
Chong et al. (2018), E. 1 0.45 [0.19, 0.72]
Johnson et al. (2003), E. 4 0.50 [0.11, 0.89]
Ramus et al. (2000), E. 2 0.51 [-0.15, 1.16]
Nazzi et al. (2000), E. 4 0.51 [0.09, 0.93]
Bosch & Sebastian-Galles (2001), E. 2 0.52 [0.13, 0.90]
Mehler et al. (1988), E. 5 0.52 [-0.28, 1.31]
Christophe & Morton (1998), E. 3 0.56 [-0.03, 1.15]
Chung (2002), E. 2 0.58 [0.25, 0.91]
Bosch & Sebastian-Galles (2001), E. 1 0.61 [0.26, 0.96]
Molnar et al. (2013), E. 1 0.63 [0.19, 1.07]
Nacar Garcia et al. (2018), E. 1 0.63 [0.11, 1.15]
Bahrick & Pickens (1988), E. 1 0.64 [-0.02, 1.29]
Bosch & Sebastian-Galles (2001), E. 2 0.66 [0.37, 0.95]
White et al. (2016), E. 1 0.66 [0.34, 0.99]
Molnar & Carreiras (2015), E. 1 0.68 [-0.13, 1.48]
Nazzi et al. (2000), E. 2 0.68 [0.09, 1.28]
Nazzi et al. (1998), E. 1 0.72 [0.22, 1.22]
Molnar et al. (2013), E. 1 0.74 [0.34, 1.15]
Molnar & Carreiras (2015), E. 1 0.78 [-0.09, 1.65]
Molnar et al. (2013), E. 1 0.79 [0.29, 1.29]
Mehler et al. (1988), E. 1 0.81 [0.02, 1.61]
Molnar & Carreiras (2015), E. 1 0.83 [0.24, 1.41]
Mehler et al. (1988), E. 4 0.86 [0.06, 1.65]
Chong et al. (2018), E. 2 0.88 [0.22, 1.54]
Bosch & Sebastian-Galles (2001), E. 1 0.92 [0.42, 1.43]
Christophe & Morton (1998), E. 1 0.93 [0.33, 1.52]
Molnar et al. (2013), E. 1 0.95 [0.18, 1.71]
Mehler et al. (1988), E. 8 0.95 [0.21, 1.69]
Mehler et al. (1988), E. 4 1.06 [0.26, 1.85]
Christophe et al. (2003), E. 1 1.06 [0.21, 1.91]
Kitamura et al. (2013), E. 6 1.09 [0.33, 1.85]
Molnar et al. (2013), E. 1 1.10 [0.44, 1.75]
Mehler et al. (1988), E. 7 1.24 [0.39, 2.10]
Mehler et al. (1988), E. 8 1.57 [0.83, 2.31]
Overall estimate 0.25 [0.14, 0.36]
-3
-2
-1
0
1
2
3
Observed Outcome

Appendix E. Forest plot for preference tasks



Appendix F. Meta-analytic model: Overall effect size

```
ESmodel = rma.mv(g_calc,
g_var_calc,
```

data = dat,
random = ~1 | study_ID/ same_infant/ experiment).
Multivariate Meta-Analysis Model (k = 160; method: REML).

logLik	Deviance	AIC	BIC	AICc
−175.8529	351.7058	359.7058	371.9814	359.9655

Variance Components:

	Estimate	sqrt	nlvls	Fixed	Factor
$\sigma^2$1	0.1060	0.3255	38	No	study_ID
$\sigma^2$2	0	0	135	No	study_ID/same_infant
$\sigma^2$3	0.0781	0.2795	138	No	study_ID/same_infant/experiment

Test for Heterogeneity:
Q(df = 159) = 810.8053, p-val < 0.0001.
Model Results:

Estimate	SE	95% CI	z	p
0.0684	0.0649	[−0.0588, 0.1955]	1.0542	0.2918

Appendix G. Meta-analytic model: Overall effect sizes by *task*

ESmodel.para = rma.mv(g_calc,
g_var_calc,
mods = ~task,
data = dat,
random = ~1 | study_ID/ same_infant/ experiment).
Multivariate Meta-Analysis Model (k = 160; method: REML).

logLik	Deviance	AIC	BIC	AICc
−163.1293	326.2586	336.2586	351.5715	336.6533

Variance Components:

	estimate	sqrt	no. levels	fixed	factor
$\sigma^2$1	0.0459	0.2143	38	no	study_ID
$\sigma^2$2	0	0	135	no	study_ID/same_infant
$\sigma^2$3	0.0731	0.2705	138	no	study_ID/same_infant/experiment

Test for Residual Heterogeneity:
QE(df = 158) = 723.3612, p-val < 0.0001.
Test of Moderators (coefficient 2):
QM(df = 1) = 30.1714, p-val < 0.0001.
Model Results:

	Estimate	SE	95% CI	z	p
Intercept	−0.0028	0.0514	[−0.1034, 0.0979]	−0.0538	0.9571
Preference-Discrimination	0.5133	0.0935	[0.3302; 0.6965]	5.4929	**<0.0001**

Appendix H. Best-fitting model: Rhythm class analysis

BestModel_RQ1 = rma.mv(g_calc,
g_var_calc,
mods = ~task/ (rhythm_class+ rhythm_class: mean_age + method),
data = dat,
random = ~1 | study_ID/same_infant/experiment).
Multivariate Meta-Analysis Model (k = 157; method: REML).

logLik	Deviance	AIC	BIC	AICc
−146.7377	293.4753	327.4753	377.7244	332.4108

Variance Components:

	estimate	sqrt	no. levels	fixed	factor
$\sigma^2$1	0.0247	0.1571	36	no	study_ID
$\sigma^2$2	0	0.0003	131	no	study_ID/same_infant
$\sigma^2$3	0.0790	0.2811	134	no	study_ID/same_infant/experiment

Test for Residual Heterogeneity:
QE(df = 142) = 623.038, p-val < 0.0001.
Test of Moderators (coefficients 2:14):
QM(df = 13) = 65.3434, p-val < 0.0001.
Model Results:

		Estimate	SE	95% CI	z	p
1	Intercept	−0.0381	0.0635	[−0.1626, 0.0864]	−0.6000	0.5485
	Task					
2	Discrimination-Preference	0.5616	0.1239	[0.3187, 0.8045]	4.5316	**<0.0001****
	Discrimination					
3	Rhythm class: Different-Same	0.2123	0.1012	[0.0139, 0.4108]	2.0970	**0.0360***
4	Same rhythm class: Mean age	−0.0399	0.0848	[−0.2061, 0.1262]	−0.4711	0.6376
5	Different rhythm class: Mean age	0.0693	0.1602	[−0.2446, 0.3832]	0.4324	0.6654
6	Method: HAS	0.1269	0.2532	[−0.3693, 0.6231]	0.5013	0.6162
7	Method: CF	0.3141	0.1293	[0.0607, 0.5675]	2.4291	**0.0151***
8	Method: EEG/NIRS	−0.5328	0.2220	[−0.9679, −0.0976]	−2.3996	**0.0164***
	Preference					
9	Rhythm class: Different-Same	−0.0614	0.1547	[−0.3646, 0.2418]	−0.3971	0.6913
10	Same rhythm class: Mean age	0.3426	0.1538	[0.0412, 0.6440]	2.2277	**0.0259***
11	Different rhythm class: Mean age	−0.1344	0.0879	[−0.3067, 0.0378]	−1.5296	0.1261
12	Method: HAS	−0.0139	0.2921	[−0.5865, 0.5587]	−0.0476	0.9621
13	Method: CF	−0.0344	0.1958	[−0.4181, 0.3493]	−0.1758	0.8605
14	Method: EEG/NIRS	0.1356	0.2110	[−0.2781, 0.5492]	0.6424	0.5206
	**LRT p < .05, ** LRT p < .01*					

Appendix I. Best-fitting model: durational metrics analysis

BestModel_RQ2 = rma.mv(g_calc,
g_var_calc,
mods = ~task/ (deltaV+ rPVI·C),
data = dat,
random = ~1 | study_ID/ same_infant/ experiment).
Multivariate Meta-Analysis Model (k = 127; method: REML).

logLik	Deviance	AIC	BIC	AICc
−129.9809	259.9619	277.9619	303.1240	279.5835

Variance Components:

	estimate	sqrt	no. levels	fixed	factor
σ^2.1	0.0397	0.1992	30	no	study_ID
σ^2.2	0	0.0000	103	no	study_ID/same_infant
σ^2.3	0.0956	0.3091	106	no	study_ID/same_infant/experiment

Test for Residual Heterogeneity:
QE(df = 121) = 578.7621, p-val < 0.0001.
Test of Moderators (coefficients 2:6):
QM(df = 5) = 39.1288, p-val < 0.0001.
Model Results:

	Estimate	SE	95% CI	z	p
Intercept	−0.0494	0.0756	[−0.1976, 0.0988]	−0.6532	0.5136
Preference-Discrimination	0.5151	0.1509	[0.2195, 0.8108]	3.4147	**0.0006****
Preference					
△V	0.0233	0.0778	[−0.1293, 0.1758]	0.2991	0.7649
rPVI-C	−0.0005	0.1060	[−0.2083, 0.2073]	−0.0049	0.9961
Discrimination					
△V	−0.1726	0.0586	[−0.2875, −0.0577]	−2.9446	**0.0032***
rPVI-C	0.1478	0.0521	[0.0457, 0.2498]	2.8382	**0.0045***

**LRT p < .05, ** LRT p < .01.*

Appendix J. Effects of all durational metrics

Meta-analytic models were run with one durational metric at a time (main effect or interaction with age) nested in task. Model output (estimates, SE, *z,* and p) are provided below.

Metric	Task	Effect	Estimate	SE	z	p
%V	discrimination	main	0.0206	0.0387	0.5313	0.5952
		:age	0.0070	0.0399	0.1743	0.8616
	preference	main	0.1283	0.1111	1.1553	0.2480
		:age	0.2730	0.1029	2.6545	**0.0079**
△C	discrimination	main	−0.0651	0.0436	−1.4938	0.1352
		:age	0.1170	0.0942	1.2422	0.2142
	preference	main	0.0337	0.0637	0.5292	0.5967
		:age	−0.1479	0.0981	−1.5082	0.1315
△V	discrimination	main	−0.1498	0.0581	−2.5779	**0.0099**
		:age	0.0570	0.0874	0.6519	0.5145
	preference	main	0.0149	0.0605	0.2465	0.8053
		:age	−0.0571	0.0502	−1.1389	0.2548
VarcoV	discrimination	main	−0.1515	0.0432	−3.5100	**0.0004**
		:age	0.0252	0.0777	0.3243	0.7457
	preference	main	0.0033	0.0717	0.0457	0.9635
		:age	−0.1151	0.0940	−1.2239	0.2210
nPVI-V	discrimination	main	−0.0345	0.0487	−0.7089	0.4784
		:age	−0.0012	0.0451	−0.0267	0.9787
	preference	main	0.0589	0.0844	0.6985	0.4848
		:age	−0.1079	0.1227	−0.8799	0.3789
rPVI-C	discrimination	main	0.1494	0.0485	3.0812	**0.0021**
		:age	0.0312	0.0605	0.5164	0.6056
	preference	main	−0.0187	0.0797	−0.2351	0.8141
		:age	0.1544	0.1144	1.3502	0.1770

Appendix K. Best-fitting model: Rhythm class and durational metrics

BestModel_Expl = rma.mv(g_calc,
g_var_calc,
mods = ~task/ (deltaV+ rPVI·C+.
rhythm_class: mean_age),
data = datsub2,
random = ~1 | study_ID/ same_infant/ experiment).
Multivariate Meta-Analysis Model (k = 125; method: REML).

logLik	Deviance	AIC	BIC	AICc
−121.0549	242.1097	272.1097	313.0205	277.0582

Variance Components:

	Estimate	sqrt	no. levels	Fixed	Factor
$\sigma^2$1	0.0179	0.1339	28	no	study_ID
$\sigma^2$2	0.0000	0.0000	101	no	study_ID/same_infant
$\sigma^2$3	0.0964	0.3105	104	no	study_ID/same_infant/experiment

Test for Residual Heterogeneity:
QE(df = 113) = 514.4825, p-val < 0.0001.
Test of Moderators (coefficients 2:6):
QM(df = 11) = 63.0785, p-val < 0.0001.
Model Results:

		Estimate	SE	95% CI	z	p
1	Intercept	−0.1047	0.0766	[−0.2548, 0.0454]	−1.3669	0.1717
	Task					
2	Discrimination-Preference	0.6568	0.1533	[0.3564, 0.9572]	4.2857	**<0.0001****
	Discrimination					
3	△V	−0.1511	0.0608	[−0.2703, −0.0319]	−2.4843	**0.0130****
4	rPVI-C	0.1499	0.0519	[0.0482, 0.2515]	2.8899	**0.0039***
5	Rhythm class	0.0469	0.1176	[−0.1835, 0.2773]	0.3989	0.6900
6	Same rhythm class: Mean age	−0.0700	0.1088	[−0.2832, 0.1431]	−0.6438	0.5197
7	Different rhythm class: Mean age	−0.1110	0.1122	[−0.3308 0.1088]	−0.9895	0.3224
	Preference					
8	△V	0.1767	0.0989	[−0.0171, 0.3705]	1.7866	0.0740
9	rPVI-C	−0.0105	0.1141	[−0.2342, 0.2131]	−0.0924	0.9263
10	Rhythm class	−0.4509	0.2686	[−0.9774, 0.0757]	−1.6783	0.0933
11	Same rhythm class: Mean age	0.4832	0.2044	[0.0827, 0.8838]	2.3646	**0.0180***
12	Different rhythm class: Mean age	−0.0637	0.0883	[−0.2367, 0.1094]	−0.7208	0.4710
	** LRT p < .05, ** LRT p < .01*					

Appendix L. Best-fitting model: exploratory analysis

BestModel_Expl = rma.mv(g_calc,
g_var_calc,
mods = ~task/ (deltaV+ rPVI.C*infant_type+ rhythm_class: mean_age),
data = dat,
random = ~1 | study_ID/ same_infant/ experiment).
Multivariate Meta-Analysis Model (k = 125; method: REML).

logLik	Deviance	AIC	BIC	AICc
−113.7184	227.4368	261.4368	307.4988	268.0174

Variance Components:

	Estimate	sqrt	no. levels	Fixed	Factor
$\sigma^2$1	0.0055	0.0741	28	no	study_ID
$\sigma^2$2	0.0184	0.1357	101	no	study_ID/same_infant
$\sigma^2$3	0.0614	0.2478	104	no	study_ID/same_infant/experiment

Test for Residual Heterogeneity:
QE(df = 111) = 426.0205, p-val < 0.0001.
Test of Moderators (coefficients 2:6):
QM(df = 13) = 94.4666, p-val < 0.0001.
Model Results:

	Estimate	SE	95% CI	z	p
Intercept	−0.0401	0.1560	[−0.3459, 0.2658]	−0.2567	0.7974
Preference-Discrimination	−0.5001	0.3121	[−1.1118, 0.1116]	−1.6025	0.1090
Preference					
△V	0.1030	0.0694	[−0.0331, 0.2391]	1.4836	0.1379
rPVI-C	−0.0362	0.1648	[−0.3592, 0.2868]	−0.2196	0.8262
Language background	−0.8307	0.2200	[−1.2618, −0.3995]	−3.7762	**0.0002**
Language background:rPVI-C	0.2567	0.1806	[−0.0974, 0.6107]	1.4210	0.1553
Same rhythm class: Mean age	0.6850	0.1815	[0.3292, 1.0408]	3.7736	**0.0002**
Different rhythm class: Mean age	−0.0323	0.0790	[−0.1872, 0.1226]	−0.4088	0.6827
Discrimination					
△V	−0.1047	0.0568	[−0.2160, 0.0066]	−1.8445	0.0651
rPVI-C	0.4523	0.1389	[0.1801, 0.7246]	3.2562	**0.0011**
Language background	0.5219	0.2455	[0.0408,1.0030]	2.162	**0.0335**
Language background:rPVI-C	−0.3265	0.1472	[−0.6150, −0.0381]	−2.2189	**0.0265**
Same rhythm class: Mean age	−0.1068	0.0963	[−0.2955, 0.0819]	−1.1088	0.2675
Different rhythm class: Mean age	−0.1505	0.0849	[−0.3169, 0.0160]	−1.7717	0.0764

References

Abboub, N., Boll-Avetisyan, N., Bhatara, A., Höhle, B., & Nazzi, T. (2016). An exploration of rhythmic grouping of speech sequences by French- and German-learning infants. *Frontiers in Human Neuroscience, 10*.

Abercrombie, D. (1967). *Elements of general phonetics*. Edinburgh University Press.

Afshartous, D., & Preston, R. A. (2011). Key results of interaction models with centering. *Journal of Statistics Education, 19*(3), 1.

Aiken, L. S., West, S. G., & Reno, R. R. (1991). *Multiple regression: Testing and interpreting interactions*. Inc: Sage Publications.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723.

Arvaniti, A. (2012). The usefulness of metrics in the quantification of speech rhythm. *Journal of Phonetics, 40*(3), 351–373.

Arvaniti, A., & Rodriquez, T. (2013). The role of rhythm class, speaking rate, and F0 in language discrimination. *Laboratory Phonology, 4*(1), 7–38.

Bahrick, L. E., & Pickens, J. N. (1988). Classification of bimodal English and Spanish language passages by infants. *Infant Behavior and Development, 11*(3), 277–296.

Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003). Do rhythm measures tell us anything about language type?. In *Proceedings of the 15th ICPhS* (pp. 2693–2696).

Benavides-Varela, S., Hochmann, J.-R., Macagno, F., Nespor, M., & Mehler, J. (2012). Newborn's brain activity signals the origin of word memories. *Proceedings of the National Academy of Sciences, 109*(44), 17908–17913.

Bergmann, C., Rabagliati, H., & Tsuji, S. (2019). *What's in a looking time preference?*.

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.

Bertoncini, J., Bijeljac-Babic, R., Jusczyk, P. W., Kennedy, L. J., & Mehler, J. (1988). An investigation of young infants' perceptual representations of speech sounds. *Journal of Experimental Psychology: General, 117*(1), 21–33.

Bird, S., Fais, L., & Werker, J. F. (2005, May). *The phonetic rhythm/syntactic headedness connection: Evidence from Tagalog [poster]. Semiannual meeting of the Acoustical Society of America*. British Columbia, Canada: Vancouver.

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th annual conference of the Cognitive Science Society (pp. 124–129)*. Cognitive Science Society.

Bloch, B. (1950). Studies in colloquial Japanese IV phonemics. *Language, 26*(1), 86.

Boll-Avetisyan, N., Omane, P. O., & Kügler, F. (2020, May 25). Speech rhythm in Ghanaian languages: The cases of Akan, ewe and Ghanaian English. In *Proceedings of the 9th international conference on speech prosody. International Conference on Speech Prosody, Tokyo, Japan*.

Borzone de Manrique, A. M., & Signorini, A. (1983). Segmental duration and rhythm in Spanish. *Journal of Phonetics, 11*(2), 117–128.

Bosch, L. (2010, April 16). *Rhythm cues and language discrimination in infancy: A review* [workshop]. Workshop on Prosodic Development, Barcelona. http://prosodia.upf.edu/activitats/prosodicdevelopment/presentacions/bosch.pdf.

Bosch, L., Cortés, C., & Sebastián-Gallés, N. (2001). El reconocimiento temprano de la lengua materna: Un estudio basado en la voz masculina [Early native-language recognition capacities: A study based on male-voices]. *Infancia y Aprendizaje, 24*(2), 197–213.

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.

Bosch, L., & Sebastián-Gallés, N. (2001). Evidence of early language discrimination abilities in infants from bilingual environments. *Infancy, 2*(1), 29–49.

Butler, J., Floccia, C., Goslin, J., & Panneton, R. (2011). Infants' discrimination of familiar and unfamiliar accents in speech. *Infancy, 16*(4), 392–417.

Butler, J., & Frota, S. (2018). Emerging word segmentation abilities in European Portuguese-learning infants: New evidence for the rhythmic unit and the edge factor. *Journal of Child Language, 45*(6), 1294–1308.

Byers-Heinlein, K., Burns, T. C., & Werker, J. F. (2010). The roots of bilingualism in newborns. *Psychological Science, 21*(3), 343–348.

Carbajal, M. J., Peperkamp, S., & Tsuji, S. (2021). A meta-analysis of infants' word-form recognition. *Infancy, 26*(3), 369–387.

Chong, A. J., Vicenik, C., & Sundara, M. (2018). Intonation plays a role in language discrimination by infants. *Infancy, 23*(6), 795–819.

Christophe, A., & Morton, J. (1998). Is Dutch native English? Linguistic analysis by 2-month-olds. *Developmental Science, 1*(2), 215–219.

Christophe, A., Nespor, M., Guasti, M. T., & van Ooijen, B. (2003). Prosodic structure and syntactic acquisition: The case of the head-direction parameter. *Developmental Science, 6*(2), 213–222.

Chung, T. (2002). *Speech accent categorization in infancy*. Doctoral dissertation. University of Pittsburgh http://d-scholarship.pitt.edu/10414/1/Ting-ting_chung_dissertation.pdf.

Cichocki, W., Selouani, S.-A., & Perreault, Y. (2014). Measuring rhythm in dialects of New Brunswick French: Is there a role for intensity? *Canadian Acoustics, 42*(3), 2.

Clopper, C. G., & Smiljanic, R. (2015). Regional variation in temporal organization in American English. *Journal of Phonetics, 49*, 1–15.

Coetzee, A. W., & Wissing, D. P. (2007). Global and local durational properties in three varieties of south African English. *The Linguistic Review, 24*(2–3).

Cristia, A. (2013). Input to language: The phonetics and perception of infant-directed speech. *Language and Linguistics Compass, 7*(3), 157–170.

Cristia, A., Minagawa-Kawai, Y., Egorova, N., Gervain, J., Filippin, L., Cabrol, D., & Dupoux, E. (2014). Neural correlates of infant accent discrimination: An fNIRS study. *Developmental Science, 17*(4), 628–635.

Cristia, A., Tsuji, S., & Bergmann, C. (2020). *Theory evaluation in the age of cumulative science*.

Dauer, R. M. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics, 11*(1), 51–62.

Dehaene-Lambertz, G., & Houston, D. (1998). Faster orientation latencies toward native language in two-month-old infants. *Language and Speech, 41*(1), 21–43.

Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for deltaC. In P. Karnowski, & I. Szigeti (Eds.), *Language and language-processing (pp. 231–241)*. Peter Lang.

Diehl, M., Varga, K., Panneton, R., Burnham, D., & Kitamura, C. (2006, June 19). Six-month-old infants' perception of native speech accent. *Annual Meeting of the XVth Biennial International Conference on Infant Studies*. XVth Biennial International Conference on Infant Studies, Westin Miyako, Kyoto, Japan. http://citation.allacademic.com/meta/p94069_index.html.

Ding, H., & Xu, X. (2016). *L2 English rhythm in read speech by Chinese students* (pp. 2696–2700).

Dunlap, W. P., Cortina, J. M., Vaslow, J. B., & Burke, M. J. (1996). Meta-analysis of experiments with matched groups or repeated measures designs. *Psychological Methods, 1*(2), 170–177.

Enzinna, N. R. (2016). Spanish-influenced rhythm in Miami English. *Proceedings of the Linguistic Society of America, 1*, 34.

Fava, E., Hull, R., & Bortfeld, H. (2014). Dissociating cortical activity during processing of native and non-native audiovisual speech from early to late infancy. *Brain Sciences, 4*(3), 471–487.

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). Irr: Various coefficients of interrater reliability and agreement. *R package version 0.84.1*. https://CRAN.R-project.org/package=irr.

Gavaldà-Ferré, N. (2007). *Vowel reduction and Catalan speech rhythm*. Unpublished Master thesis. University College.

Gervain, J. (2018). The role of prenatal experience in language development. *Current Opinion in Behavioral Sciences, 21*, 62–67.

Gervain, J., & Mehler, J. (2010). Speech perception and language Acquisition in the First Year of life. *Annual Review of Psychology, 61*(1), 191–218.

Gleitman, L. R., & Wanner, E. (1982). *Language acquisition: The state of the art*. CUP Archive.

Grabe, E., & Low, E. L. (2002). Durational variability in speech and the rhythm class hypothesis. *Papers in Laboratory Phonology, 7*, 515–546.

Grenon, I., & White, L. (2008). Acquiring rhythm: A comparison of L1 and L2 speakers of Canadian English and Japanese. In *Proceedings of the 32nd Boston University Conference on Language Development* (pp. 155–166).

Hagmann, L., & Dellwo, V. (2014). Listeners may rely on intonation to distinguish languages of different rhythm classes. *Loquens, 1*(1), Article e008.

Hayashi, A., Tamekawa, Y., & Kiritani, S. (2001). Developmental change in auditory preferences for speech stimuli in Japanese infants. *Journal of Speech, Language, and Hearing Research, 44*(6), 1189–1200.

Hedges, L. V. (1981). Distribution theory for Glass's estimator of effect size and related estimators. *Journal of Educational Statistics, 6*(2), 107–128.

Höhle, B., Bijeljac-Babic, R., & Nazzi, T. (2020). Variability and stability in early language acquisition: Comparing monolingual and bilingual infants' speech perception and word recognition. *Bilingualism: Language and Cognition, 23*(1), 56–71.

Houston-Price, C., & Nakai, S. (2004). Distinguishing novelty and familiarity effects in infant preference procedures. *Infant and Child Development, 13*(4), 341–348.

Johnson, E. K., & Braun, B. (2011). *The role of intonation in language discrimination by 4.5-month-olds*. Montreal, Canada: Society for Research in Child Development.

Johnson, E. K., Jusczyk, P. W., & Ramus, F. (2003). The role of segmental information in language discrimination by English-learning 5-month-olds. In D. Houston, A. Seidl, G. Hollich, E. K. Johnson, & A. Jusczyk (Eds.), *Jusczyk Final Report*. http://hincapie.psych.purdue.edu/Jusczyk/pdf/Segmental.pdf.

Kaminskaïa, S. (2020). *Rythme prosodique et variation stylistique en français canadien [Prosodic rhythm and stylistic variation in Canadian French]*. Congrès de l'ACL Université Western, University of Waterloo. https://incl.pl/affiches-2020-posters/2020-ACL-affiche-Kaminskaia.pdf.

Kawase, S., Kim, J., & Davis, C. (2016). *The influence of second language experience on Japanese-accented English rhythm* (pp. 746–750).

Kinzler, K. D., Dupoux, E., & Spelke, E. S. (2007). The native language of social cognition. *Proceedings of the National Academy of Sciences, 104*(30), 12577–12580.

Kitamura, C., Panneton, R., & Best, C. T. (2013). The development of language constancy: Attention to native versus non-native accents. *Child Development, 84*(5), 1686–1700.

Kitamura, C., Panneton, R., Diehl, M., & Notley, A. (2006). Attuning to the native dialect: When more means less. In *Proceedings of the 11th Australian international conference on Speech Science & Technology. Australian International Conference on Speech Science & Technology*. University of Auckland, New Zealand.

Ladefoged, P. (1975). *A course in phonetics*. Harcourt Brace Jovanovich.

Langus, A., Mehler, J., & Nespor, M. (2017). Rhythm in language acquisition. *Neuroscience & Biobehavioral Reviews, 81*, 158–166.

Lee, C. S., Kitamura, C., Burnham, D., & Todd, N. P. (2014). On the rhythm of infant-versus adult-directed speech in Australian English. *Journal of the Acoustical Society of America, 136*(1), 357–365.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., … Moher, D. (2009). The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: Explanation and elaboration. *PLoS Medicine, 6*(7), Article e1000100.

Lidji, P., Palmer, C., Peretz, I., & Morningstar, M. (2011). Listeners feel the beat: Entrainment to English and French speech rhythms. *Psychonomic Bulletin & Review, 18*(6), 1035–1041.

Lin, H., & Wang, Q. (2007). Mandarin rhythm: An acoustic study. *Journal of Chinese Language and Computing, 17*(3), 127–140.

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis*. Sage Publications, Inc.

Loukina, A., Rosner, B., Kochanski, G., Keane, E., & Shih, C. (2013). What determines duration-based rhythm measures: Text or speaker? *Laboratory Phonology, 4*(2).

Mairano, P. (2011). *Rhythm typology: Acoustic and perceptive studies*. Doctoral dissertation. Università degli Studi di Torino https://tel.archives-ouvertes.fr/tel-00654261/document.

Mairano, P., & Romano, A. (2011). Rhythm metrics for 21 languages. In *17th International Congress of Phonetic Sciences* (pp. 1318–1321).

Mairano, P., & Romano, A. (2011b). Rhythm metrics on syllables and feet do not work as expected. In *Twelfth Annual Conference of the International Speech Communication Association* (pp. 1857–1860). https://www.researchgate.net/publication/221479879_Rhythm_Metrics_on_Syllables_and_Feet_do_not_Work_as_Expected.

May, L., Byers-Heinlein, K., Gervain, J., & Werker, J. F. (2011). Language and the newborn brain: Does prenatal language experience shape the neonate neural response to speech? *Frontiers in Psychology, 2*.

Mehler, J., Dupoux, E., Nazzi, T., & Dehaene-Lambertz, G. (1996). Coping with linguistic diversity: The infant's viewpoint. In J. L. Morgan, & K. Demuth (Eds.), *Signal to syntax: Bootstrapping from speech to grammar in early acquisition* (pp. 101–116). L. Erlbaum Associates.

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

Minagawa-Kawai, Y., van der Lely, H., Ramus, F., Sato, Y., Mazuka, R., & Dupoux, E. (2011). Optical brain imaging reveals general auditory and language-specific processing in early infant development. *Cerebral Cortex, 21*(2), 254–261.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLoS Medicine, 6*(7), Article e1000097.

Mok, P. P. K. (2009). On the syllable-timing of Cantonese and Beijing mandarin. *Chinese Journal of Phonetics, 2*, 148–154.

Molnar, M., & Carreiras, M. (2015, June). Language preferences of monolingual infants from bilingual and monolingual communities. In *Workshop of infant language development (WILD 2015)*. Sweden: Stockholm University.

Molnar, M., Carreiras, M., & Gervain, J. (2016). Language dominance shapes non-linguistic rhythmic grouping in bilinguals. *Cognition, 152*, 150–159.

Molnar, M., Gervain, J., & Carreiras, M. (2013). Within-rhythm class native language discrimination abilities of Basque-Spanish monolingual and bilingual infants at 3.5 months of age. *Infancy, 19*(3), 326–337.

Moon, C., Cooper, R. P., & Fifer, W. P. (1993). Two-day-olds prefer their native language. *Infant Behavior and Development, 16*(4), 495–500.

Moon, C., Lagercrantz, H., & Kuhl, P. K. (2013). Language experienced *in utero* affects vowel perception after birth: A two-country study. *Acta Paediatrica, 102*(2), 156–160.

Morgan, J. L., & Demuth, K. (Eds.). (1996). *Signal to syntax: Bootstrapping from speech to grammar in early acquisition*. L. Erlbaum Associates.

Mukai, Y., & Tucker, B. V. (2015). Rhythm metrics of spontaneous speech and accent. *Canadian Acoustics, 43*(3), 2.

Nácar Garcia, L., Guerrero-Mosquera, C., Colomer, M., & Sebastián-Gallés, N. (2018). Evoked and oscillatory EEG activity differentiates language discrimination in young monolingual and bilingual infants. *Scientific Reports, 8*(1), 2770.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology. Human Perception and Performance, 24*, 756–766.

Nazzi, T., & Cutler, A. (2019). How consonants and vowels shape spoken-language recognition. *Annual Review of Linguistics, 5*(1), 25–47.

Nazzi, T., Iakimova, G., Bertoncini, J., Fredonie, S., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language, 54*(3), 283–299.

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., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication, 41*(1), 233–243.

Nespor, M., Peña, M., & Mehler, J. (2003). On the different roles of vowels and consonants in speech processing and language acquisition. *Lingue Linguaggio, 2*, 203–230.

Nespor, M., Shukla, M., & Mehler, J. (2011). Stress-timed vs. syllable-timed languages. In M. van Oostendorp, C. J. Ewen, E. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (pp. 1–13). John Wiley & Sons, Ltd.

Nolan, F., & Jeon, H.-S. (2014). Speech rhythm: A metaphor? *Philosophical Transactions of the Royal Society B: Biological Sciences, 369*(1658), 20130396.

Oakes, L. M. (2010). Using habituation of looking time to assess mental processes in infancy. *Journal of Cognition and Development, 11*(3), 255–268.

Paillereau, N., Podlipský, V. J., Smolík, F., Šimáčková, Š., & Chládková, K. (2021). The development of infants' sensitivity to native versus non-native rhythm. *Infancy*.

Paquette-Smith, M., & Johnson, E. K. (2015). Spanish-accented English is Spanish to English-learning 5-month-olds. In The Scottish Consortium for ICPhS 2015 (Ed.), *Vol. 18. Proceedings of the 18th International Congress of Phonetic Sciences*. The University of Glasgow. https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2015/Papers/ICPHS0262.pdf.

Peña, M., Pittaluga, E., & Mehler, J. (2010). Language acquisition in premature and full-term infants. *Proceedings of the National Academy of Sciences, 107*(8), 3823–3828.

Pettorino, M., & Pellegrino, E. (2016). %V and VtoV: An acoustic-perceptual approach to the rhythmic classification of languages. In C. Bardel, & A. De Meo (Eds.), *Parler les langues romanes; Parlare le lingue romanze; Hablar las lenguas romances; Falando línguas românicas* (pp. 13–28). Il Torcoliere. http://opar.unior.it/1926/1/GSCP_2014.pdf.

Phan, J., & Houston, D. M. (2009). Infant dialect discrimination. *The Journal of the Acoustical Society of America, 125*(4), 2776.

Pike, K. L. (1945). *The intonation of American English*. University of Michigan Press.

Pons, F., Bosch, L., & Lewkowicz, D. J. (2015). Bilingualism modulates infants' selective attention to the mouth of a talking face. *Psychological Science, 26*(4), 490–498.

Prieto, P., Vanrell, M., Astruc, L., Payne, E., & Post, B. (2012). Phonotactic and phrasal properties of speech rhythm. Evidence from Catalan, English, and Spanish. *Speech Communication, 54*(6), 681–702.

R Core Team. (2020). *R: A language and environment for statistical computing (3.6.3)*. R Foundation for Statistical Computing. https://www.Rproject.org/.

Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2019). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science, 22*(1), Article e12704.

Ramus, F. (2002). Language discrimination by newborns: Teasing apart phonotactic, rhythmic, and intonational cues. *Annual Review of Language Acquisition, 2*, 85–115.

Ramus, F., Hauser, M. D., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and by cotton-top tamarin monkeys. *Science, 288*(5464), 349–351.

Ramus, F., Nespor, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition, 73*, 265–292.

Roach, P. (1982). On the distinction between "stress-timed" and "syllable-timed" languages. In D. Crystal (Ed.), *Linguistic controversies* (pp. 73–79). Edward Arnold.

Romano, A. (2010). Speech rhythm and timing: Structural properties andacoustic correlates. In S. Schmid, M. Schwarzenbach, & D. Studer (Eds.), *La dimensione temporale del parlato* (pp. 45–75). EDK Editore. https://www.semanticscholar.org/paper/Speech-Rhythm-and-Timing%3A-Structural-Properties-and-Romano/4fdeba58260f517bb51050ee6489da23ac1a190b.

Roy, J.-P., Macoir, J., Martel-Sauvageau, V., & Boudreault, C.-A. (2012). Two French-speaking cases of foreign accent syndrome: An acoustic–phonetic analysis. *Clinical Linguistics & Phonetics, 26*(11–12), 934–945.

RStudio Team. (2020). *RStudio: Integrated development environment for R (1.3.959) [computer software]*. RStudio, PBC. http://www.rstudio.com/.

de Ruiter, K., Geambasu, A., & Levelt, C. (2015). *Testing the possibility of artificial bilingualism: On the effect of non-live exposure to a foreign language on language preference in infants*. unpublished bachelor thesis. Universiteit Leiden.

Sato, H., Hirabayashi, Y., Tsubokura, H., Kanai, M., Ashida, T., Konishi, I., Uchida-Ota, M., Konishi, Y., & Maki, A. (2012). Cerebral hemodynamics in newborn infants exposed to speech sounds: A whole-head optical topography study. *Human Brain Mapping, 33*(9), 2092–2103.

Shafer, V. L., Shucard, D. W., & Jaeger, J. J. (1999). Electrophysiological indices of cerebral specialization and the role of prosody in language acquisition in 3-month-old infants. *Developmental Neuropsychology, 15*(1), 73–109.

Shousterman, C. (2014). Speaking English in Spanish Harlem: The role of rhythm. *University of Pennsylvania Working Papers in Linguistics, 20*(2), 159–168.

Soderstrom, M., Xu Rattanasone, N., Demuth, K., & Seidl, A. ((unpublished). *Unpublished data)*. 2021.

Sterne, J. A. C., Sutton, A. J., Ioannidis, J. P. A., Terrin, N., Jones, D. R., Lau, J., … Higgins, J. P. T. (2011). Recommendations for examining and interpreting funnel plot asymmetry in meta-analyses of randomised controlled trials. *BMJ, 343*(jul22 1), d4002.

Stojanovic, D. (2013). *Cross-linguistic comparison of rhythmic and phonotactic similarity [doctoral dissertation]*. University of Hawaii.

Toledo, G. (2009). Métricas rítmicas en tres dialectos Amper-Hispanoamérica. *Ianua. Revista Philologica Romanica, 9*, 1–21.

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-augmented meta-analyses: Toward cumulative data assessment. *Perspectives on Psychological Science, 9*(6), 661–665.

Turk, A., & Shattuck-Hufnagel, S. (2013). What is speech rhythm? A commentary on Arvaniti and Rodriquez, Krivokapić, and Goswami and Leong. *Laboratory Phonology, 4*(1).

Venables, W. N., & Ripley, B. D. (2002). *Modern applied statistics with S* (4th ed.). Springer http://www.stats.ox.ac.uk/pub/MASS4.

Verhoeven, J. (2020). *The north wind and the sun in phonetics*. Phonetics Expert. https://www.phonetics.expert/north-wind-and-the-sun.

Vicenik, C. (2011). *The role of intonation in language discrimination by infants and adults*. Doctoral dissertation. University of California http://phonetics.linguistics.ucla.edu/research/Vicenik_Diss.pdf.

Vicenik, C., & Sundara, M. (2013). The role of intonation in language and dialect discrimination by adults. *Journal of Phonetics, 41*(5), 297–306.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3).

Vieru, B., de Mareüil, P. B., & Adda-Decker, M. (2011). Characterisation and identification of non-native French accents. *Speech Communication, 53*(3), 292–310.

Weissenborn, J., & Höhle, B. (Eds.). (2001), *vol. 2. Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition*. John Benjamins Publishing.

Wenk, B. J., & Wioland, F. (1982). Is French really syllable-timed? *Journal of Phonetics, 10*(2), 193–216.

White, L., Benavides-Varela, S., & Mády, K. (2020). Are initial-consonant lengthening and final-vowel lengthening both universal word segmentation cues? *Journal of Phonetics, 81*, 100982.

White, L., Floccia, C., Goslin, J., & Butler, J. (2014). Utterance-final lengthening is predictive of infants' discrimination of English accents: Infants' accent discrimination. *Language Learning, 64*(s2), 27–44.

White, L., Luche, C. D., & Floccia, C. (2016). Five-month-old infants' discrimination of unfamiliar languages does not accord with "rhythm class". In *Proceedings of Speech Prosody* (pp. 567–571).

White, L., & Malisz, Z. (2020). Speech rhythm and timing. In C. Gussenhoven, & A. Chen (Eds.), *The Oxford Handbook of Language Prosody* (pp. 165–180). Oxford University Press.

White, L., & Mattys, S. L. (2007). Calibrating rhythm: First language and second language studies. *Journal of Phonetics, 35*(4), 501–522.

White, L., Mattys, S. L., & Wiget, L. (2012). Language categorization by adults is based on sensitivity to durational cues, not rhythm class. *Journal of Memory and Language, 66*(4), 665–679.

White, L., Payne, E., & Mattys, S. L. (2009). Rhythmic and prosodic contrast in Venetan and Sicilian Italian. In M. Vigário, S. Frota, & M. J. Freitas (Eds.), *Current Issues in Linguistic Theory (Vol. 306, pp. 137–158)*. John Benjamins Publishing Company.

Wickham, H. (2016). *ggplot2: Elegant graphics for data analysis*. Springer-Verlag. https://ggplot2.tidyverse.org.

Zacharaki, K., & Sebastián-Gallés, N. (2019). Language discrimination abilities of 4–5 month old monolingual and bilingual infants. In *4th Workshop on Infant Language Development (WILD 2019)*. Germany: University of Potsdam. https://www.uni-potsdam.de/wild2019/programme.html.

Zacharaki, K., & Sebastian-Galles, N. (2021). The ontogeny of early language discrimination: Beyond rhythm. *Cognition, 104628*.