Article:
Cox, Christopher, Keren-Portnoy, Tamar orcid.org/0000-0002-7258-2404, Roepstorff, Andreas et al. (1 more author) (2022) A Bayesian Meta-Analysis of Infants' Ability to Perceive Audio-VisualCongruence for Speech. Infancy. pp. 67-96. ISSN: 1532-7078

https://doi.org/10.1111/infa.12436

A Bayesian Meta-Analysis of Infants' Ability to Perceive Audio-Visual Congruence for Speech

Christopher Martin Mikkelsen Cox [1, 2, 3] (ccox@cc.au.dk)
Tamar Keren-Portnoy [3] (tamar.keren-portnoy@york.ac.uk)
Andreas Roepstorff [1, 2] (andreas.roepstorff@cas.au.dk)
Riccardo Fusaroli [1, 2] (fusaroli@cc.au.dk)

2021

[1] School of Communication and Culture, Aarhus University
Jens Chr. Skous Vej 2, Building 1485, 8200 Aarhus, Denmark

[2] Interacting Minds Centre, Aarhus University
Jens Chr. Skous Vej 4, Building 1483, 8200 Aarhus, Denmark

[3] Department of Language and Linguistic Science, University of York
V/C/210, 2nd Floor, Block C, Vanbrugh College,
Heslington, York YO10 5 DD, United Kingdom

Correspondence: chris.mm.cox@gmail.com (Chris Cox)

Keywords: Bayesian meta-analysis, cognitive development, multimodal integration, audio-visual matching

Word Count: 8096 words

The data and code for this meta-analysis are available on:
https://github.com/4CCoxAU/MA_audiovisual_congruence

A Bayesian Meta-Analysis of Infants' Ability to Perceive Audio-Visual Congruence for Speech

0.0: Abstract

This paper quantifies the extent to which infants can perceive audio-visual congruence for speech information and assesses whether this ability changes with native-language exposure over time. A hierarchical Bayesian robust regression model of 92 separate effect sizes extracted from 24 studies indicates a moderate effect size in a positive direction (0.35, CI [0.21: 0.50]). This result suggests that infants possess a robust ability to detect audio-visual congruence for speech. Moderator analyses, moreover, suggest that infants' audio-visual matching ability for speech emerges at an early point in the process of language acquisition and remains stable for both native and non-native speech throughout early development. A sensitivity analysis of the meta-analytic data, however, indicates that a moderate publication bias for significant results could shift the lower credible interval to include null effects. Based on these findings, we outline recommendations for new lines of enquiry and suggest ways to improve the replicability of results in future investigations.

1.0: Introduction

There is more to a face than meets the eye. Infants not only see faces, but also hear them. Human faces and voices suffuse infants' perceptual experience from birth and play a crucial role in their social and emotional development. Infants thus experience a multisensory world that requires integration of time-locked information across modalities. This form of intermodal integration may play a significant role within the domain of language acquisition, where temporally coupled auditory and visual information emanates from the faces of speakers. By combining cues about speech sounds from multiple modalities, infants can obtain information beyond what would be possible if they were relying on independent estimates from individual senses. This synchronous flow of perceptual cues from multiple modalities raises questions about the extent to which infants can integrate this information during language development and whether linguistic exposure engenders change in this ability over the course of development. By delving deeper into the multimodality of infants' language acquisition process, this allows for a more ecological construal of how infants discover and acquire the speech sounds of their ambient language.

1.1: The Building Blocks of Perceptual Development

Over the course of early development, infants learn how to parse a dynamic world of changing multimodal information with little experience and limited attentional resources to guide them. Infants may learn to derive structure from this flow of multimodal information by attending to

constant patterns across variation in input (Bremner, Lewkowicz, & Spence, 2012; Bahrick & Lickliter, 2012). For example, the temporal synchronisation of salient speech information across the auditory and visual modalities may direct infant attention towards this multimodal relationship when exposed to a speaking face. The fundamental skills required for this intersensory matching ability appear to develop early in infancy (Bahrick & Lickliter, 2000; Lewkowicz, 2000, 2014). For example, 5-month-old infants are shown to be able to detect changes in the rhythm of a toy hammer striking a wooden surface when the rhythm is presented bimodally, but not when it is presented unimodally (Bahrick & Lickliter, 2000). Infants have also been shown to exhibit flexibility in their combination of multisensory cues and can take advantage of the intersensory redundancy of audio-visual speech stimuli to adapt to the difficulty of the task at hand (Lewkowicz and Hansen-Tift, 2012; Pons, Bosch, and Lewkowicz, 2015; Hillairet de Boisferon et al., 2017). For example, 12-month-old infants exposed to non-native speech have been shown to revert back to the attentional allocation patterns characteristic of younger infants; that is, the infants attend more to the mouth region in order to facilitate their processing of the non-native speech stream (Lewkowicz and Hansen-Tift, 2012; Hillairet de Boisferon et al., 2017). Infants thus show sophistication in their ability to integrate independent speech cues across modalities, as also exemplified by studies relying on the McGurk effect (Burnham & Dodd, 2004; Kushnerenko, Teinonen, Volein, & Csibra, 2008; Rosenblum, Schmuckler, & Johnson, 1997; Desjardins & Werker, 2004). These studies indicate that the low-level congruence of salient events across modalities functions as a building block of perceptual development and allows infants to parse coherent multimodal events using a unified perceptual system (Gibson, 1969; Bahrick & Lickliter, 2012; Lewkowicz, 2000, 2014). This meta-analysis focuses broadly on the extent to which infants can detect audio-visual congruence for speech information (i.e., infants' ability to perceive correspondences between speech cues across modalities) and how this ability changes over the course of development, as explained further below.

1.2: The Experimental Paradigm

Studies investigating the development of infants' ability to perceive intersensory congruence examine whether infants can associate stimuli via two different sensory channels (Féron, Gentaz, & Streri, 2006; Filippetti, Johnson, Lloyd-Fox, Dragovic, & Farroni, 2013; Sann & Streri, 2007). A coherent subset of these studies investigates infants' ability to perceive audio-visual congruence for speech, defined here as a correspondence between visual speech (e.g. a visual display of a face with spread or rounded lips) and auditory speech (e.g. an auditory token [i] or [u], etc.). The experimental paradigm involves presenting two side-by-side visual displays of faces producing a

speech sound together with an auditory stimulus that is congruent with only one of the faces. The purpose of these studies is to investigate the extent to which infants can attend to the audio-visual stimuli that match across the sensory modalities. Kuhl and Meltzoff (1982), who were some of the first to establish this cross-modal matching procedure, show that infants as young as 4.5 months look significantly longer towards the face that matches the heard vowel. These differential looking responses suggest that infants perceive the congruence between visual and auditory speech information at an early point in language development.

1.3: Intersensory Matching Ability Appears Early in Development

The body of research amassed since Kuhl and Meltzoff's (1982) first study replicates and extends the above results, but suggests a more complex pattern of development. For example, experimental studies show that neonates exhibit sensitivity to audio-visual congruence (Aldridge, Braga, Walton, & Bower, 1999; Coulon, Hemimou, & Streri, 2013) and that infants under the age of three days exhibit longer and faster looking times towards audio-visual congruence than incongruence (Guellai, Streri, Chopin, Rider, & Kitamura, 2016). Although these studies use slightly modified experimental procedures (e.g. infant-controlled trials and simplified visual stimuli) to accommodate neonate attentional limitations, these findings indicate that infants may be able to perceive audio-visual concordance with minimal experience. The early onset of this ability suggests that infants detect multimodal congruence by initially relying on the low-level synchrony of salient events across the different modalities (cf. Lewkowicz, 2010; Bahrick & Lickliter, 2000; Lewkowicz et al., 2015).

1.4: Perceptual Narrowing according to Language Familiarity

These early perceptual abilities, however, may decline for non-native languages over the course of development. Pons, Lewkowicz, Soto-Faraco, and Sebastián-Gallés (2009) show that 6-month-old Spanish-learning infants can perceive audio-visual congruence for both native and non-native consonants (/ba/ vs. /va/), while 11-month-old Spanish-learning infants perform at chance level for the non-native speech sound. This pattern of perceptual specialisation finds further support in studies showing that 10-14-month-old English-learning infants can detect audio-visual congruence in fluent passages of native speech more reliably and faster than in passages of non-native speech (Lewkowicz, Minar, Tift, & Brandon, 2015). The authors of these studies claim that as infants gain experience with the auditory and visual cues of their native language phonology, their ability to detect multisensory coherence in non-native phonemic contrasts declines. This form of perceptual reorganisation according to the properties of infants' ambient language is well-attested by studies

involving auditory-only stimuli (e.g. Kuhl et al., 2006; Segal, Hejli-Assi, & Kishon-Rabin, 2016; for a review, see Werker & Gervain, 2013), and taken together with the results from studies on audio-visual congruence, these findings may be reflective of a common developmental mechanism that mediates perceptual narrowing effects (cf. Lewkowicz & Ghazanfar, 2006). In the following meta-analysis, we analyse how language familiarity interacts with infants' ability to perceive audio-visual congruence over the course of early infancy and discuss these developmental patterns further in section 4.1.1. It should be noted that analysing language familiarity as a binary moderator variable (i.e. native vs. non-native) disregards the potential influence of the phonological status and acoustic distinctiveness of speech sounds in the respective languages, as discussed further in section 4.2.1.

1.5: Intersensory Matching Exhibits Complex Patterns of Development

Other experimental findings indicate complex patterns of development. For example, 12-month-old German-learning infants can only detect congruence in a non-native language, while younger infants are able do so in both their native language and a non-native language (Kubicek, Hillairet de Boisferon, et al., 2014). This developmental pattern appears to interact with speech style in intricate ways; for example, 12-month-old German-learning infants are shown to perceive intersensory coherence for their native language only if sentences are spoken in an infant-directed speech style (Kubicek, Gervain, et al., 2014), whereas 6-month-old French-learning infants gaze significantly longer to congruence for adult-directed speech than for infant-directed speech (Richoz et al., 2017). Other experimental studies further obscure the developmental patterns by showing that 4-month-old infants' audio-visual matching performance can be disrupted by conflicting gender information (Patterson & Werker, 2002), and that 4.5-month-old infants' ability to perceive congruence depends on subtle inter-speaker differences in the visual distinctiveness of their vowel articulation (Pejovic, Yee, & Molnar, 2020). Further studies admit a role for stimulus complexity affecting infants' ability to perceive audio-visual congruence; whereas Lewkowicz, Minar, Tift, and Brandon (2015) show that infants aged 12 and 14 months, but not infants aged 4, 8 and 10 months, can match fluent passages of auditory speech to synchronous faces, Kubicek et al. (2014) and Dorn, Weinert, and Falck-Ytter (2018) show that infants as young as 4.5 months can perceive audio-visual congruence for fluent speech in both native and non-native speech. In the following meta-analysis, we therefore investigate stimulus complexity as a moderator in order to clarify whether infants' ability to perceive audio-visual congruence depends on the complexity of the stimuli (i.e. fluent passages of speech versus simple syllables) and whether this ability changes over the course of development.

1.6: The Concurrent Development of Intersensory Matching in the Motor Domain

Infants' ability to perceive audio-visual congruence may also be influenced by the concurrent development of intersensory matching in the motor domain. Many of the above authors note that infants make mouth movements that match those of the speakers in experimental trials (Coulon et al., 2013; Kuhl & Meltzoff, 1982; Kuhl & Meltzoff, 1996; Legerstee, 1990). The notion that motor aspects of speech production play a role in infants' audio-visual matching ability is examined directly by Yeung and Werker (2013) who show that 4.5-month-old infants exhibit selective impairment of audio-visual matching if the articulatory movements relevant for the production of a specific speech sound are restricted by teethers. This finding on the sensorimotor underpinnings of infants' ability to perceive audio-visual congruence receives further support in longitudinal studies showing that the ability correlates with vocal productivity across development (Altvater-Mackensen, Mani, & Grossmann, 2016; Streri, Coulon, Marie, & Yeung, 2016). This close connection between infants' ability to perceive audio-visual congruence and their concurrent motor development in early infancy may contribute to the results of this meta-analysis and warrants further study, as discussed further in section 4.1.2 below.

2.0: Aims

The above experimental findings fail to produce a straightforward pattern of development for infants' ability to detect audio-visual congruence for speech information during the process of language acquisition. The following meta-analysis quantifies the robustness of this ability and assesses the current evidence for whether it changes with native-language exposure over the course of early infancy. The aggregation of results across studies permits investigation of the extent to which the following five moderator variables influence infants' audio-visual matching ability: i) age, ii) language familiarity, iii) the interaction between age and language familiarity, iv) stimulus complexity, v) the interaction between age and stimulus complexity. The justification for each will be described in brief: i) Firstly, it is of interest to establish whether infants' ability to detect audio-visual congruence changes over the course of early infancy. By pooling together data from the meta-analytic studies, we can examine whether infants' capability undergoes developmental change or remains stable throughout early development. This moderator variable also allows us to investigate whether age produces a shift in infants' preference for audio-visual congruence. Studies show that infants' preferences for novelty relate to infant age as well as stimulus complexity (Hunter & Ames, 1988; Rose, Gottfried, Melloy-Carminar & Bridger, 1982; Kidd, Aslin & Piantadosi, 2012; 2014). Most of the above studies show that infants exhibit longer looking times towards audio-visual congruence for speech information (e.g. Kuhl and Meltzoff, 1982; Guellai et

al., 2016; Pons et al., 2009; Lewkowicz et al., 2015), but other studies indicate that infants prefer to attend to audio-visual incongruence at various points in development (e.g. Streri et al., 2016; Pejovic, Yee, & Molnar, 2020). By including age as a moderator variable, we can explore whether infants initially prefer to attend to audio-visual congruence (i.e. a familiarity response) and thereafter shift to seek out audio-visual incongruence (i.e. a novelty response) as they gain more experience with the multisensory contingencies of speech. This latter pattern of development would manifest as a shift from a positive effect size to a negative effect size. ii) Secondly, the extent to which infants' audio-visual matching ability undergoes perceptual narrowing still remains an open question, and we examine language familiarity as a moderator variable to explore whether this capacity differs for native versus non-native stimuli. If this variable moderates infants' ability to detect audio-visual congruence, this would imply that its development depends on infants' experience with the auditory and visual cues of their native language phonology. iii) Thirdly, we investigate the interaction between age and language familiarity as a moderator in order to explore whether infants exhibit differential response patterns to audio-visual congruence in native and non-native stimuli over the course of early development. If infants' experience with the specific audio-visual co-occurrences in their native language mediates their audio-visual matching ability (i.e. if perceptual narrowing applies to infants' multisensory perception), then this would manifest as a decline in infants' ability to detect audio-visual congruence for non-native speech stimuli over time. iv) Fourthly, we examine the complexity of the stimuli as a moderator variable in order to compare infants' proportion of looking time towards auditory stimuli comprised of individual speech segments versus fluent passages. v) Relatedly, infants' ability to perceive audio-visual congruence for fluent passages of speech may change with age, so we examine the interaction between stimulus complexity and age as a moderator. With a view to assessing the evidence for this audio-visual matching capacity and formulating recommendations that can inform future investigations, the following meta-analysis aims to determine the heterogeneity between studies, to calculate the magnitude of the pooled effect size, and to examine the potential influence of moderators.

2.1: Methodology

In order to obtain a comprehensive set of peer-reviewed results, we adopted the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Guidelines (PRISMA, Stewart et al., 2015), as shown in Appendix A below, and conducted a systematic literature search on PubMed, Web of Science, and Google Scholar using the following combination of search terms: *(cross-modal OR audio-visual OR intermodal OR multimodal) AND (matching OR congruence or concordance) AND (speech*

perception OR perception) AND (infant OR toddler* OR infancy)*. By performing forward and backward literature searches in the papers identified by this initial search, 11 additional studies were identified. This search strategy yielded a total of 189 papers. These papers were then screened for inclusion according to the following criteria: i) because we are interested in the early development of infants' audio-visual matching ability and how this relates to native-language exposure, participants had to be typically-developing and aged between 0 and 15 months, ii) experiments had to involve the presentation of visual and auditory speech stimuli, and iii) the dependent measure had to be within-participant looking times. Of the initial 189 papers, 34 were duplicates, 19 papers examined non-typical infant populations, 5 papers did not examine infants in the relevant age range, 24 papers used a different methodology or measure, and 83 papers were unrelated to the topic under investigation. Because papers often contained several experiments yielding multiple effect sizes (e.g. with different age groups and speech stimuli), the final sample encompassed 24 papers and 92 individual measures of effect sizes.

By extracting the reference lists of the included studies from Web of Science using the R package *bibliometrix* (Aria & Cuccurullo, 2017), we built a network model of co-citation coupling in order to visualise relevant clusters in the literature, as shown in Fig. 1 below. The above network of co-citation coupling visualises links between papers that are cited together (i.e. co-citation) and papers that cite the same papers (i.e. coupling). The studies appear to cluster together according to the topic under investigation; the blue cluster examines the effects of gender and speaker identity on infants' intermodal matching ability (Bahrick, Hernandez-Reif, & Flom, 2005; Bahrick, Netto, & Hernandez-Reif, 1998; Hillairet de Boisferon et al., 2015; Patterson & Werker, 1999; Pickens et al., 1994; Poulindubois, Serbin, Kenyon, & Derbyshire, 1994; Richoz et al., 2017; Walker-Andrews, Bahrick, Raglioni, & Diaz, 1991), whereas the red cluster explores the perceptual narrowing of infants' intersensory matching ability (Dorn, Weinert, & Falck-Ytter, 2018; Guellai et al., 2016; Kubicek et al., 2013; Kubicek, Hillairet de Boisferon, et al., 2014; Kubicek, Gervain, et al., 2014; Lewkowicz et al., 2015; Lewkowicz & Pons, 2013; Pejovic et al., 2020; Pons et al., 2009) and its relation to speech production (Altvater-Mackensen et al., 2016; Streri et al., 2016). The green cluster includes the first studies to establish the intermodal matching procedure (Aldridge et al., 1999; Kuhl & Meltzoff, 1984; MacKain, Studdert-Kennedy, Spieker, & Stern, 1983; Patterson & Werker, 1999). In order to further explore the extent of internal co-citation among the studies, we computed the ratio between the number of actual and potential local citations according to publication date. This analysis shows that each of the studies is on average cited by 30% of the following studies, and moreover, suggests that Lewkowicz et al. (2015), Patterson and Werker (1999), Pons et al. (2009), and Kuhl and Meltzoff (1984) represent influential studies in the

literature, with 83%, 81%, 75% and 68% of subsequent studies citing them, respectively. The influence of these studies is also manifested in the below direct-citation network in Fig. 1, where the above influential studies function as central anchor points for the individual direct-citation clusters. The collection of studies under investigation, then, represents a diverse intersection of coherent clusters of experiments that examine a variety of relevant aspects of infants' ability to perceive audio-visual congruence for speech sounds.

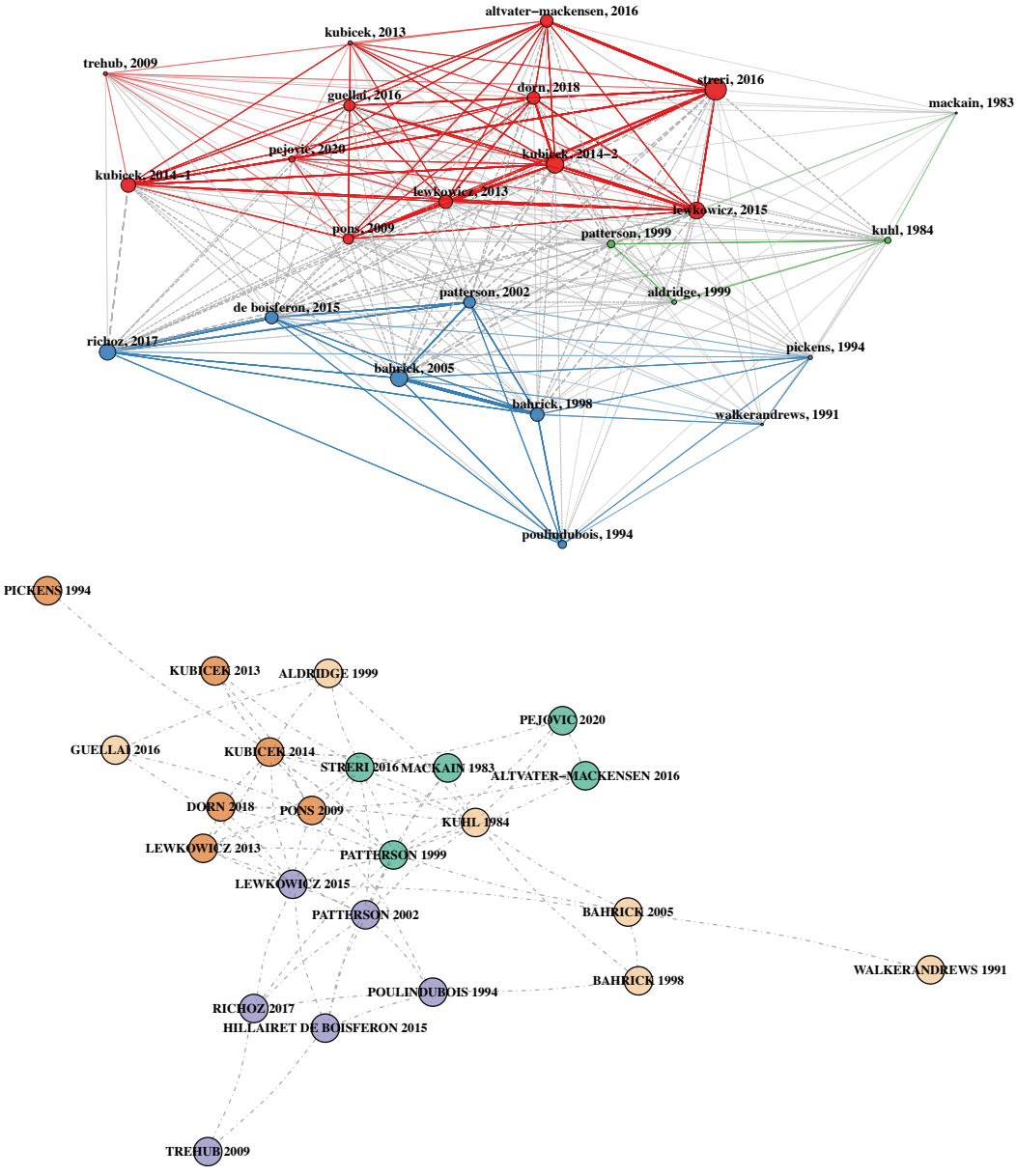


Figure 1: The above network shows the co-citation coupling strength (i.e. the number of times two studies are cited together by a new article as well as bibliographic similarity) for the final sample of cited studies. The colour and thickness of the lines represent clusters of strong citation links. The below direct-citation network shows which studies cite each other. The colours represent clusters of strong direct-citation links.

2.2: Data Extraction

This meta-analysis extracts data from published empirical studies and conforms to the ethical guidelines for conducting research with human subjects outlined in the Declaration of Helsinki. The method used to extract data from the studies depended on the reported statistics. Most of the studies expressed infants' ability to perceive audio-visual congruence in terms of the proportion of total looking time towards the voice-matched face (derived by dividing the time spent looking to the voice-matched face by the time spent looking at both faces), whereas others reported raw looking times. In order to standardise these measures and to allow for comparison between studies, we calculated Hedges' *g*, a variant of Cohen's *d* that is preferred for small sample sizes (Morris, 2000). The effect size represents the standardized mean difference between infants' looking times towards audio-visual congruence versus incongruence; the bigger the effect size, the larger the standardized mean difference. The method of computation of Hedges' *g* depended on the reported results. Some of the studies (N = 33 experiments) measured infants' baseline preference for the visual stimuli by exposing infants to silent video clips. In this case, we used standard formulae based on means and standard deviations for effect size calculations: $\text{Hedges'g} = \frac{\text{MeanLT}_1 - \text{BaselineMeanLT}_2}{\text{SD}_{\text{pooled}}}$. Other studies (N = 23 experiments) compare infants' preference for audio-visual congruence to a baseline condition of 0.50. The standard deviation of this baseline condition was estimated using binomial approximation based on the number of experimental test trials, *n* : $\sigma = \sqrt{np(1 - p)}$. For the remaining studies that did not report raw looking times, effect sizes were calculated using the reported *d*-values (N = 20 experiments) and one-sample or paired *t*-values (N = 16 experiments).

When using the *t*- and *d*-values to compute effect sizes, the standard deviation of the effect size could not be computed. In order to include these effect sizes in the meta-analysis, these missing standard deviation values were imputed by using multivariate imputation by chained equations based on a Bayesian linear regression in the R package *mice* (Groothuis-Oudshoorn & Van Buuren, 2011). In order to account for the statistical uncertainty involved in the partially stochastic process of imputation (cf. Azur, Stuart, Frangakis, & Leaf, 2011; Sterne et al., 2009), 20 datasets were constructed with sample size, mean age, and hedges' *g* values as predictors. The standard deviation values of the imputed datasets were checked for similarity to the reported standard deviations and post-processed to include only positive values. This process of multiple imputation does not appear to bias the estimation of the overall effect size, as will be explored further in Section 3.1. All hierarchical Bayesian models in this paper pool the results of analyses performed on these 20 imputed datasets. The raw data are available on MetaLab.

2.3: Meta-Analytic Model

Meta-analyses provide a pooled estimate of the overall effect size by combining the weighted results of comparable individual studies. The use of a random-effects model enables us to estimate and adjust for heterogeneity across studies and therefore to account for heterogeneity in population samples and methodologies (cf. Fernández-Castilla et al., 2020). The multi-level structure of the random-effects model posits that the true effect size may be study-specific (e.g. due to differences in study design or population) and thereby permits explicit modelling of heterogeneity in the results. The credible interval of the pooled estimate is thus a function of both within-study sampling error and between-study variance (Hedges & Vevea, 1998). This hierarchical structure serves to incorporate the correlation among multiple within-study effect sizes, the disregard of which can create undue certainty in the estimates (Fernández-Castilla et al., 2020).

In order to estimate the pooled effect size and credible intervals, a hierarchical Bayesian robust regression model using a Student's t-likelihood was fitted to the data. Robust regression methods implement longer-tailed distributions (here, a student's *t*-distribution) in order to dampen the influence of outliers, thus incorporating outliers without allowing them to dominate non-outlier data (Jylänki, Vanhatalo, & Vehtari, 2011). Weakly informative priors were chosen, so that their influence on the meta-analytic estimates were small and extreme effect sizes were discounted as unlikely (cf. Lemoine, 2019; Gelman, Simpson & Betancourt, 2017). For the overall effect, a normal distribution with a mean of 0 and standard deviation of 0.5 was chosen based on our prior expectations for effect sizes (cf. Cohen, 1988). This prior implies that we expect approximately 95% of the effect size distribution to be between -1 and 1. For the heterogeneity of the effects (i.e., the standard deviation of random effects), a positive truncated normal distribution with a mean of 0 and standard deviation of 0.2 was chosen. For the degrees of freedom parameter, ν, of the Student's t-distribution, a gamma distribution with a shape parameter of 2 and a scale parameter of 0.1 was chosen in order to ensure that the model remains robust to the influence of outliers (cf. McElreath, 2020; Kruschke, 2015). Prior predictive checks were performed to ensure that model predictions for plausible values of effect sizes would only exclude implausibly high or low values on the basis of the priors (cf. Gelman et al., 2020).

The models were fitted using Hamiltonian Monte Carlo samplers with 2 parallel chains with 20,000 iterations each, an adapt delta of 0.99 and a maximum tree depth of 20 in order to ensure no divergence in the estimation process. The quality of the models was assessed by i) performing prior and posterior predictive checks (cf. Fig. 7-8 in Appendix B), ii) ensuring Rhat statistics to be lower than 1.1, iii) plotting prior against posterior estimates and assessing whether the posteriors had lower variance than the priors (cf. Fig. 9-12 in Appendix C), iv) ensuring no divergences in

the process of estimation, and v) checking that the number of effective bulk and tail samples was above 200. To assess the extent to which the imputation of standard deviations affected the estimates, we compared the estimate of the meta-analytic effect size without imputation to that with imputation, as reported in section 3.1.

Because the dataset includes experiments that explicitly pit factors of infants' detection of audio-visual congruence against each other, this may be expected to lead to different effect sizes within the same dataset. Based on the questions raised in the above literature review and in order to analyse the influence of potential moderators on the variation of effect sizes across studies, we fit five separate models to the meta-analytic data with the following predictors: i) age modelled as a monotonic non-linear function, ii) language familiarity, iii) stimulus complexity, iv) an interaction between age and language familiarity, and v) an interaction between age and stimulus complexity. Age was modelled as a monotonic non-linear function in order to examine whether the relative direction of developmental change in effect sizes is modulated by age, without assuming that this potential change occurs at a constant rate. Leave-one-out (loo) information criteria and stacking weights were calculated in order to asses which of the models had the lowest pointwise out-of-sample prediction accuracy and generalised best to new data (Vehtari, Gelman, & Gabry, 2017).

Publication bias was assessed by conducting quantitative sensitivity analyses and by inspecting a significance funnel plot, following the methods introduced by Mathur & VanderWeele (2020). These methods assume that meta-analytic studies represent samples from an underlying population of published and unpublished studies, where the probability of selection for significant studies is higher. The potential presence of publication bias is thereby assessed by estimating how much more likely significant studies are to be published than non-significant studies and by calculating the amount of publication bias required to attenuate the point estimate or its credible interval to a given value.

All computations were performed in R 4.0 (R Core Team, 2013) using *brms* 2.14 (Bürkner, 2017) and *Stan* 2.21 (Carpenter et al., 2017) in RStudio 1.2 (RStudio Team, 2018). The analysis scripts are available on github: https://github.com/4CCoxAU/MA_audiovisual_congruence and osf: https://osf.io/yx68a/ .

3.0: Results

3.1: Effect Size Estimate

The pooled effect size based on 92 measures of Hedges' *g* from 24 studies (cf. Fig. 2) reveals an overall estimate of 0.349 with 95% CI [0.205, 0.503], with a between-study variance of 0.26 [0.11, 0.43] and within-study variance of 0.12 [0.00, 0.29]. According to Cohen's (1988) criteria for effect

size estimates, this pooled result indicates a small to medium effect. A standardised mean difference of this size implies that approximately 60% of infants' looking times will be longer towards audio-visual congruence than incongruence.

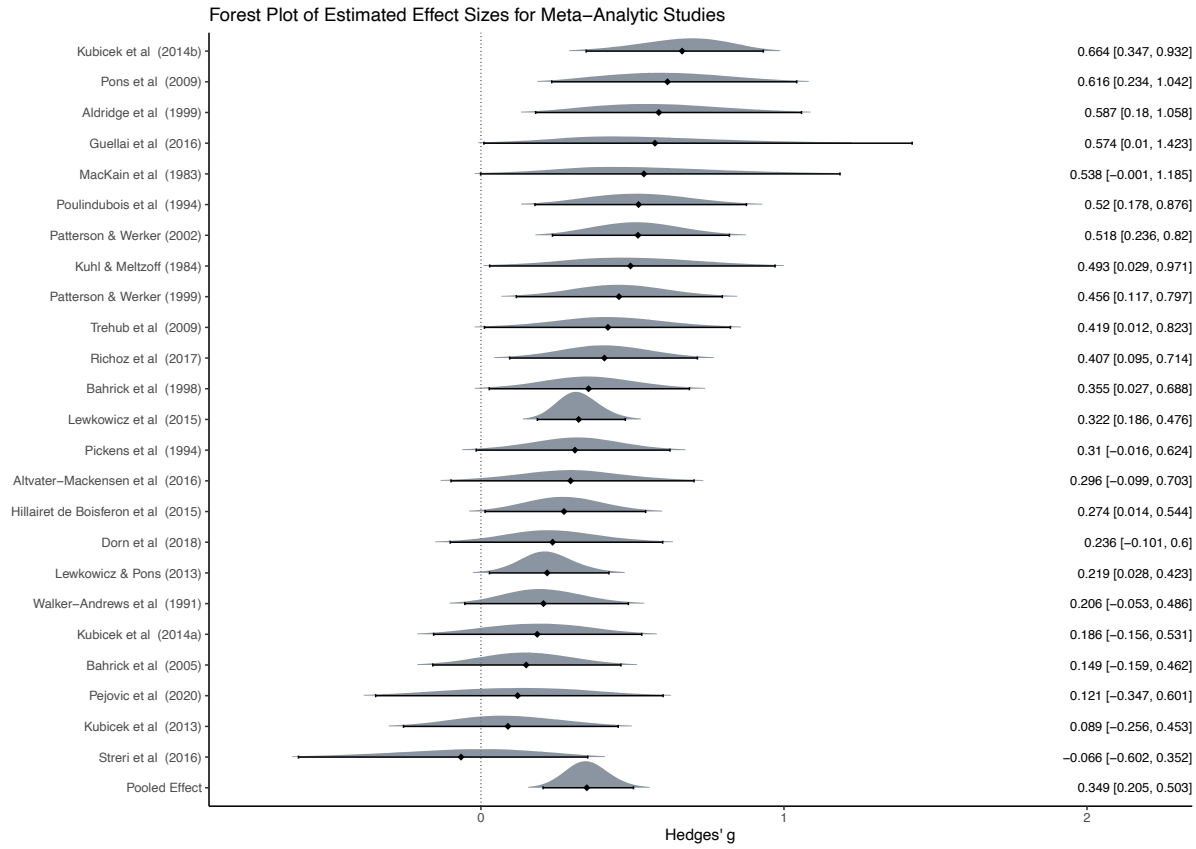


Figure 2: Estimated effect sizes for infants' detection of audio-visual congruence across experimental studies. The shaded areas indicate the posterior probability density of each estimate. The numbers to the right provide the estimated mean effect size (Hedges' *g*) and upper and lower 95% credible intervals. See Fig. 18 in Appendix F for a forest plot of effect size estimates according to the individual experiments.

To assess the extent to which the imputation of standard deviations affected the meta-analytic estimates, we compared the estimate of the meta-analytic effect size without imputation (0.24 with 95% CI [0.03, 0.46], with 56 measures) to that with imputation (0.35 with 95% CI [0.21, 0.50], with 92 measures). As the effect size estimate with the imputed datasets lies within the credible interval of the non-imputed dataset, and since most of the studies without standard deviation are centred on positive values, there does not appear to be evidence of bias on the estimation of the overall effect size as a result of the data imputation process.

3.2: Moderator Analysis for Heterogeneity

The leave-one-out cross-validation performed on the baseline model showed that model generalisability did not increase by adding the following moderators: age modelled as a monotonic non-linear function (stacking weight = 0.00), language familiarity (stacking weight = 0.00), the interaction between age and language familiarity (stacking weight = 0.00), stimulus complexity (stacking weight = 0.00), and the interaction between age and stimulus complexity (stacking weight = 0.00). These results parallel those obtained from visual inspection of a plot with the conditional effects of age and language familiarity, as shown below in Fig. 3. The posterior predictions from the meta-analytic model demonstrate that the overall direction of change in the effect size can be both positive and negative. This pattern suggests no systematic trajectory of development for infants' ability to perceive audio-visual congruence over the course of early infancy. The strong effect sizes observed for infants under the age of 50 days, moreover, may be a product of the slightly modified experimental procedures used to accommodate young infants' attentional limitations (cf. Aldridge et al., 1999; Guellai et al., 2016). These analyses suggest that infants' ability to perceive audio-visual congruence remains stable and unaffected by age, language familiarity and stimulus complexity over the course of early development (cf. Fig. 13-16 in Appendix D). The implications and limitations of these results will be discussed further in Section 4.1 below.

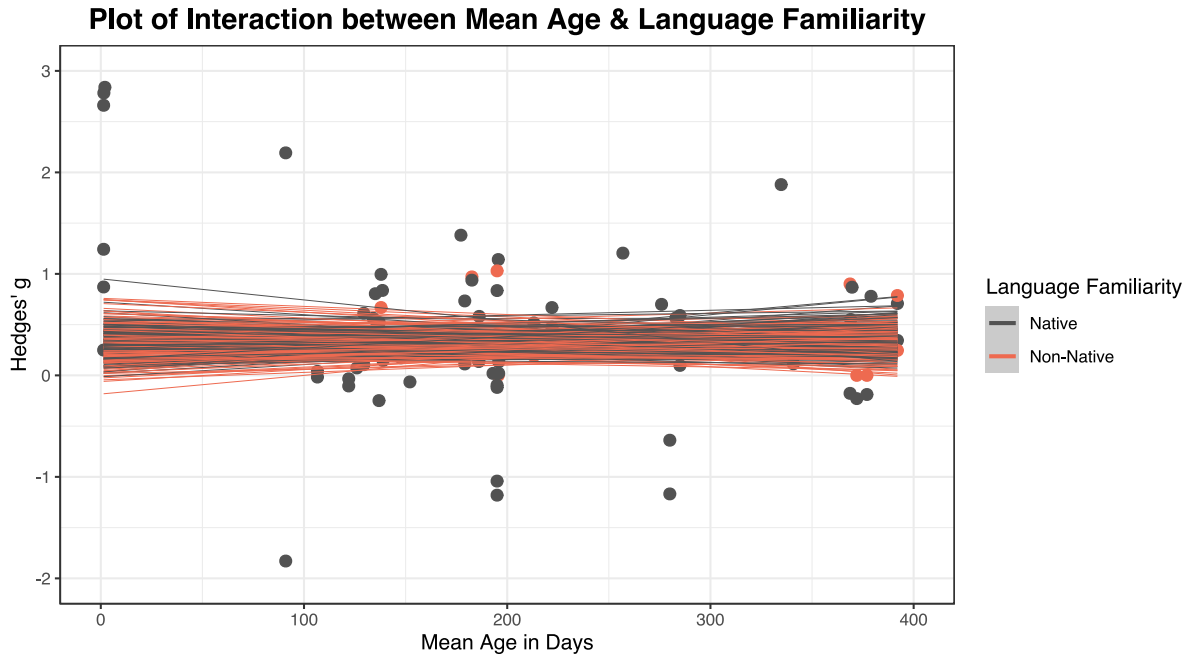


Figure 3: Spaghetti plot showing 150 posterior model predictions for the interaction between age and language familiarity. The data indicate no clear developmental patterns in neither a positive nor negative direction. The effect size estimates, however, remain above zero, which indicates that this ability remains stable over the course of early infancy.

3.3: Publication Bias

3.3.1: Sensitivity Analysis & Significance Funnel Plot

By treating the publication probability of significant studies as an unknown sensitivity parameter, we can estimate the severity of the publication bias required to attenuate the credible interval of the pooled effect size to include values below a specific threshold (Mathur & VanderWeele, 2020). A quantitative sensitivity analysis with a random-effects specification indicates that significant results would need to be at least 1.48 times more likely to be published for the credible interval to include an effect size of 0.10, at least 2-fold more likely to be published to include an effect size of 0.05, and at least 3-fold more likely to include an effect size of 0, as depicted in Fig. 4 below. These estimates of publication bias severity represent relatively plausible values, thus suggesting that moderate publication bias would be required to shift the lower credible interval to include null effects (cf. Mathur & VanderWeele, 2020).

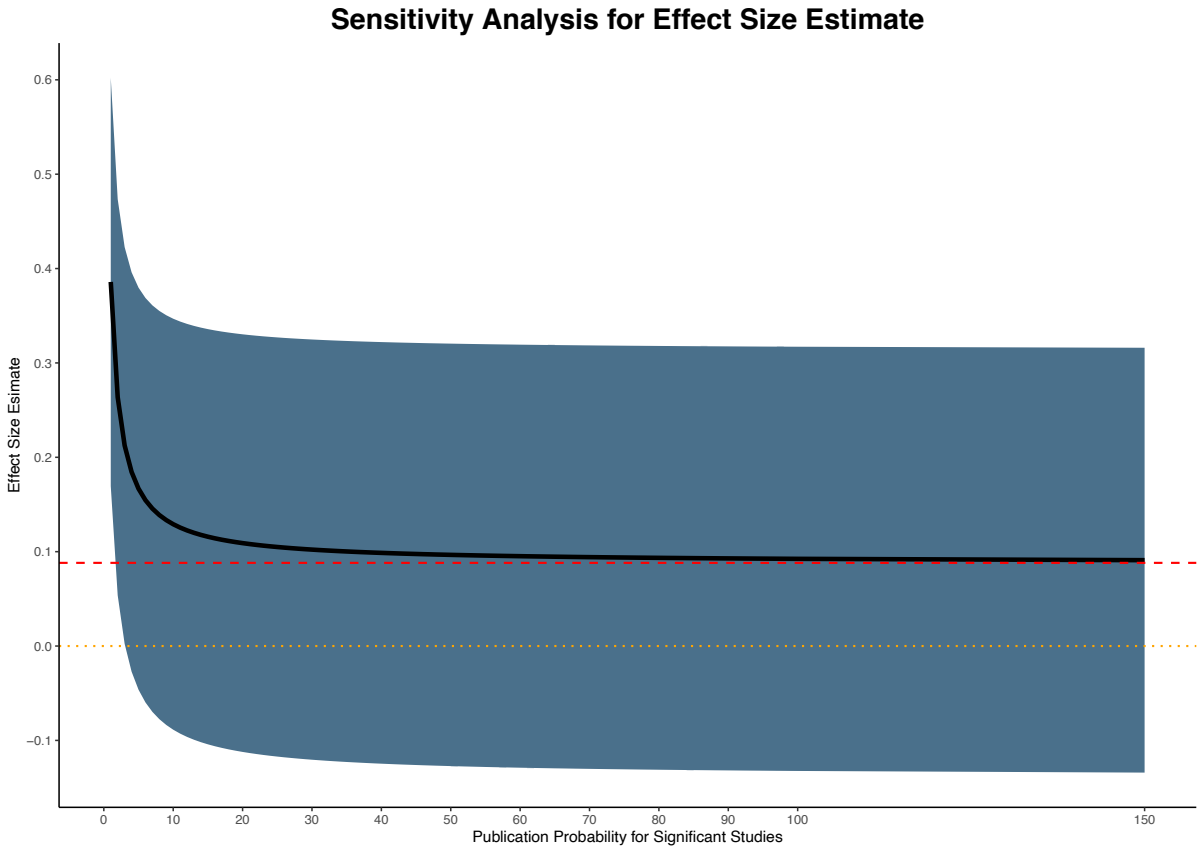


Figure 4: Plot of sensitivity analysis showing the effect size estimate as a function of severity of publication bias. The plot indicates what happens to the effect size if the publication probability is x times higher for significant studies than for non-significant studies. An effect size estimate of 0.0 is indicated by the orange dotted line, and the worst-case point estimate (see below) is indicated by the dashed red line.

3.3.2: Significance Funnel Plot

By using the inverse of the sum of study variance and a heterogeneity estimate, we can compute the uncorrected worst-case estimate for the effect size based solely on non-significant studies: 0.088 with 95% CI [-0.139, 0.316], as plotted in the significance funnel plot in Fig. 5 below. Although the significance funnel plot indicates a weak correlation between the point estimates and their squared standard errors, the estimates for non-significant studies primarily revolve around the null. This distributional pattern further implies that the meta-analytic estimate is fragile to moderate and quite plausible publication bias.

Together with the results of the above sensitivity analysis, these findings indicate that our estimates are likely inflated, were a realistic publication bias for significant findings present in the literature (cf. Tsuji, Cristia, Frank, & Bergmann, 2019; Von Holzen & Bergmann, 2018). Based on these results, the estimates should be interpreted with caution, and we outline suggestions to improve the replicability and transparency of results in future investigations in section 4.2 below.

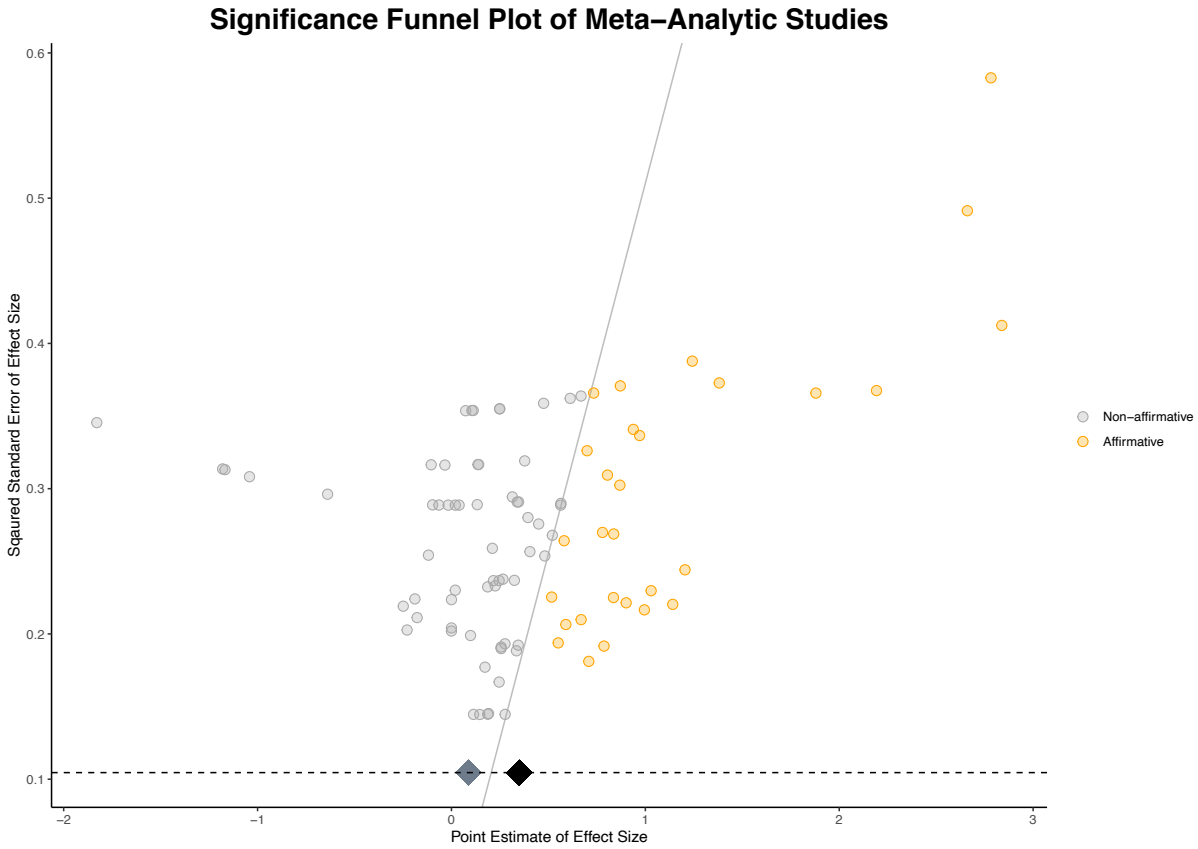


Figure 5: Significance funnel plot of studies. Studies on the diagonal line have exactly p = 0.05. Grey diamond: worst-case estimate of effect size based only on non-significant studies. Black diamond: estimate of effect size for all studies. The significance funnel plot indicates a weak correlation between point estimates and their standard errors.

4.0: Discussion

This paper set out to review and quantify evidence for infants' ability to perceive audio-visual congruence for speech sounds in order to formulate recommendations that can inform future investigations. The aggregation of data from 24 experimental studies suggested a modest effect size in a positive direction, indicating that infants prefer to attend to audio-visual congruence for speech sounds. The above citation network analyses showed that the studies under investigation examine a wide variety of relevant aspects of infants' ability to detect audio-visual congruence for speech. Based on these network analyses and the reviewed literature, we investigated whether infants exhibit a developmental shift in their attentional preference and the extent to which this audio-visual matching capacity undergoes perceptual narrowing. The experimental data provided no evidence for a developmental change in infants' preference to attend to audio-visual congruence, nor for a reliable effect of language familiarity or stimulus complexity. These results suggest that infants can perceive intermodal congruence for both native and non-native speech at an early point in language acquisition and that this audio-visual matching ability remains stable over the course of development (cf. Fig. 3), as discussed further below. Moreover, the quantitative analyses of publication bias sensitivity suggest that were even a moderate bias towards significant results present in the literature, the underlying effect may be closer to null. The above findings warrant consideration of several issues in future investigations of infants' ability to perceive audio-visual congruence.

4.1: Implications & Directions for Future Research

4.1.1: The Stability of Infants' Audio-Visual Matching Ability

Both the absence of a shift in infants' attentional preference for audio-visual congruence and the lack of evidence for perceptual narrowing may be a product of the multisensory redundancy inherent in audio-visual speech stimuli. Experimental studies on how infants' intersensory matching abilities develop over the course of early infancy show that infants can flexibly adapt their allocation of attention to useful aspects of the multimodal information stream. Lewkowicz and Hansen-Tift (2012), for example, use an eye-tracking methodology to show that infants exhibit striking developmental shifts in their selective attention towards facial regions in response to audio-visual speech stimuli. Whereas 4-month-old infants attend significantly more to the eye region of caregivers and 6-month-old infants look equally to caregivers' eyes and mouths, 8- and 10-month-old infants attend significantly more to the mouth region. Because the mouth region provides the most reliable and salient visual cues to speech, infants' allocation of attention to this facial area at the onset of babble production (i.e. 8 to 10 months) may serve to facilitate their

processing of the speech signal. This explanation for the developmental shift in attentional allocation receives further support from Lewkowicz and Hansen-Tift's (2012) evidence on 12-month-old infants; at this age, the infants are shown no longer to attend to the mouth region in response to native audio-visual speech stimuli, but they revert back to doing so when exposed to non-native audio-visual speech stimuli. Other studies obtain similar results to the patterns in this study (Hillairet de Boisferon et al., 2017) and extend the findings by showing that bilingual infants overall attend more to the visual mouth region than monolingual infants, presumably to better disambiguate the two languages (Pons, Bosch, and Lewkowicz, 2015). This set of results implies that infants can flexibly take advantage of the intersensory redundancy of audio-visual speech stimuli and adapt to the difficulty of the task at hand.

The sophistication of infants' combination of multisensory cues for speech finds further support in studies relying on the McGurk effect, an illusory perception effect that arises due to the incongruence between auditory and visual speech input (e.g. when an auditory [ba] is presented concurrently with an incongruent visual [ga]). When faced with this incongruent multimodal information, adults have been shown to perceive a fusion between the auditory and visual stimuli, resulting in a percept of [da] (McGurk & MacDonald, 1976). Several studies employing both behavioural and electro-physiological paradigms suggest that 4- and 5-month-old infants show similar effects (Burnham & Dodd, 2004; Kushnerenko, Teinonen, Volein, & Csibra, 2008; Rosenblum, Schmuckler, & Johnson, 1997; Desjardins & Werker, 2004). These studies indicate that infants possess a powerful capacity to integrate independent speech cues across modalities.

The meta-analytic finding of no perceptual narrowing in infants' ability to detect audio-visual congruence, and the failure to produce evidence of a straightforward pattern of perceptual narrowing in empirical studies (cf. Pons, Lewkowicz, Soto-Faraco, & Sebastián-Gallés, 2009; Lewkowicz, Minar, Tift, & Brandon, 2015; Kubicek, Hillairet de Boisferon, et al., 2014; Kubicek, Gervain, et al., 2014), then, may arise due to infants' ability to flexibly profit from salient attributes in the non-native audio-visual speech signal; that is, to overcome the greater difficulty of processing non-native speech stimuli, infants may flexibly revert back to a reliance on redundant visual information in order to ensure successful processing of the speech signal. This form of compensatory cue combination may underlie the complex developmental patterns in the literature and admits an important role for synchronicity in infants' ability to match audio-visual information (cf. Pons et al., 2009; Lewkowicz et al., 2015). As this finding of no perceptual narrowing (although note the limitations of this meta-analysis in section 4.2.1) stands in notable contrast to well-attested patterns of perceptual narrowing in unimodal domains (cf. Werker & Gervain, 2013; Maurer & Werker, 2014), this warrants further fine-grained empirical study on how infants flexibly adapt

their allocation of attention to multimodal stimuli in response to varying degrees of task difficulty and stimulus synchronicity.

4.1.2: The Relation of Audio-Visual Congruence to the Motor Domain

While this meta-analysis provides some evidence that infants can perceive the congruence between speech sound cues in multiple modalities, the relation of this capacity to the concurrent development of intersensory matching in the motor domain remains underexplored. This idea has a long history. Meltzoff and Kuhl (1994), for example, suggest that young infants' early babbling may develop their auditory-articulatory connections for speech and help them in matching visual articulations with auditory sounds. As noted in the literature review, there is strong evidence that infants' ability to detect audio-visual congruence is influenced by the concurrent development of intersensory matching in the motor domain (Kuhl & Meltzoff, 1982, 1984; Patterson & Werker, 1999; Yeung and Werker, 2013). Electrophysiological studies, moreover, show evidence that sensorimotor and multisensory processing areas in the brain co-activate (Hickok, Houde, & Rong, 2011; Dick, Solodkin, & Small, 2010) and demonstrate that infant motor areas exhibit activation in both auditory and audio-visual speech tasks (Bristow et al., 2008; Imada et al., 2006), especially for non-native speech processing (Kuhl, Ramirez, Bosseler, Lin, & Imada, 2014). Although these studies do not show that motor information is strictly required for audio-visual matching (cf. Matchin, Groulx, & Hickok, 2014), they do suggest a crucial role for the concurrent development of intersensory matching in the motor domain. As only two of the studies in our final sample included information about infants' level of production practice, however, this aspect of infants' ability to perceive audio-visual congruence remained outside the scope of the current meta-analysis. The findings from these two longitudinal studies suggest that infants' preference to attend to audio-visual congruence is modulated by the specific speech sounds in their own production repertoire (Altvater-Mackensen et al., 2016; Streri et al., 2016). These results parallel developmental patterns in similar studies on auditory-only speech perception, where vocal production initiates shifts in how infants attend to auditory stimuli (DePaolis, Vihman, & Keren-Portnoy, 2011; DePaolis, Vihman, & Nakai, 2013; Majorano, Vihman, & DePaolis, 2014). Although the current meta-analysis found no evidence for a developmental shift in infants' preference to attend to audio-visual congruence, this analysis could not take fine-grained production measures into account. The intimate connection between infants' motor development and multimodal integration - and how this association becomes bootstrapped in early infancy - warrants further study in order to provide a more complete characterisation of the role of audio-visual matching in language development.

4.2: Recommendations for Future Research

4.2.1: Limitations

Meta-analyses exhibit limitations, one of which is that they are only as good as the data they contain. Because this meta-analysis has a broad focus on infants' ability to perceive audio-visual congruence for speech, it pools information from a diverse intersection of experiments examining a variety of relevant aspects, as noted in our citation network analyses. There are certain limitations associated with this; for example, the analysis of language familiarity (i.e. native versus non-native) as a binary moderator variable neglects the potential influence of differences in the acoustic distinctiveness and phonological status of speech sounds in the respective languages. For example, results from experiments investigating German infants' ability to perceive audio-visual congruence for passages of fluent speech in Swedish (Dorn, Weinart, & Falck-Ytter, 2018) cannot be considered equivalent to experiments investigating Spanish infants' ability to perceive audio-visual congruence for English /ba/ and /va/ syllables (Pons et al., 2009). The meta-analytic conclusion that there is no effect of neither language familiarity nor for an interaction between language familiarity and age on infants' ability to perceive audio-visual congruence, then, may be premature given the above limitations and the limited age range of the studies devoted to this area of investigation (cf. Figure 3 above and Figure 17 in Appendix E). In this context, it should be mentioned that age as a moderator also represents a slightly simplified analysis of infants' developmental patterns. Studies across various domains, for example, show that older infants can revert back to the attentional allocation patterns characteristic of younger infants when task complexity or cognitive load are increased (Berger, 2004; Kidd, Piantadosi, & Aslin, 2012). Despite the above limitations, however, it should be noted that all of the above studies conform to the requirements laid out in our selection criteria and investigate relevant aspects of the extent to which infants can perceive audio-visual congruence for speech. The hierarchical structure and random-effects specification of the meta-analytic model, moreover, serves to account for study heterogeneity in experimental stimuli and methodologies and to provide a robust pooled estimate on infants' broad ability to detect audio-visual congruence. With a view to allowing researchers to test specific hypotheses with subsets of the meta-analytic studies, we have made the data and our code for this project available on the open MetaLab repository.

4.2.2: Sample Size Calculations

There are several points to consider when planning future studies on infants' ability to perceive audio-visual congruence. Based on the pooled effect size estimate, researchers would have to test minimum 66 infants to achieve 80% power in a one-sample t-test against chance level (calculated

with the R package *pwr* (Champely, 2016)). Because the median sample size per study in the above dataset is 31 participants, this implies a 46.9% probability of finding a significant result, and in turn, suggests that 53.1% of attempts to replicate these findings should fail. If we base sample size decisions on the worst-case effect size point estimate of 0.088, as calculated above, the sample size required to achieve a power of 80% in a one-sample t-test would be at least 1015 participants (or at least 799 infants for a one-tailed one-sample t-test). It should be noted, however, that almost all of the above studies employ repeated-measures designs, which can reduce intra-subject variability and thereby increase statistical power (Guo, Logan, Glueck, & Muller, 2013). If we assume a correlation among within-subject repeated-measures of 0.5 in a one-way analysis of variance, based on the pooled effect size estimate researchers would have to test minimum 41 infants in five trials or 27 infants in ten trials in order to achieve 80% power. Based on the worst-case effect size point estimate of 0.088, researchers would have to test at least 618 infants in five trials or 406 infants in ten trials in order to achieve 80% power. Figure 6 shows how using repeated-measures designs can reduce the sample size required to achieve sufficient statistical power.

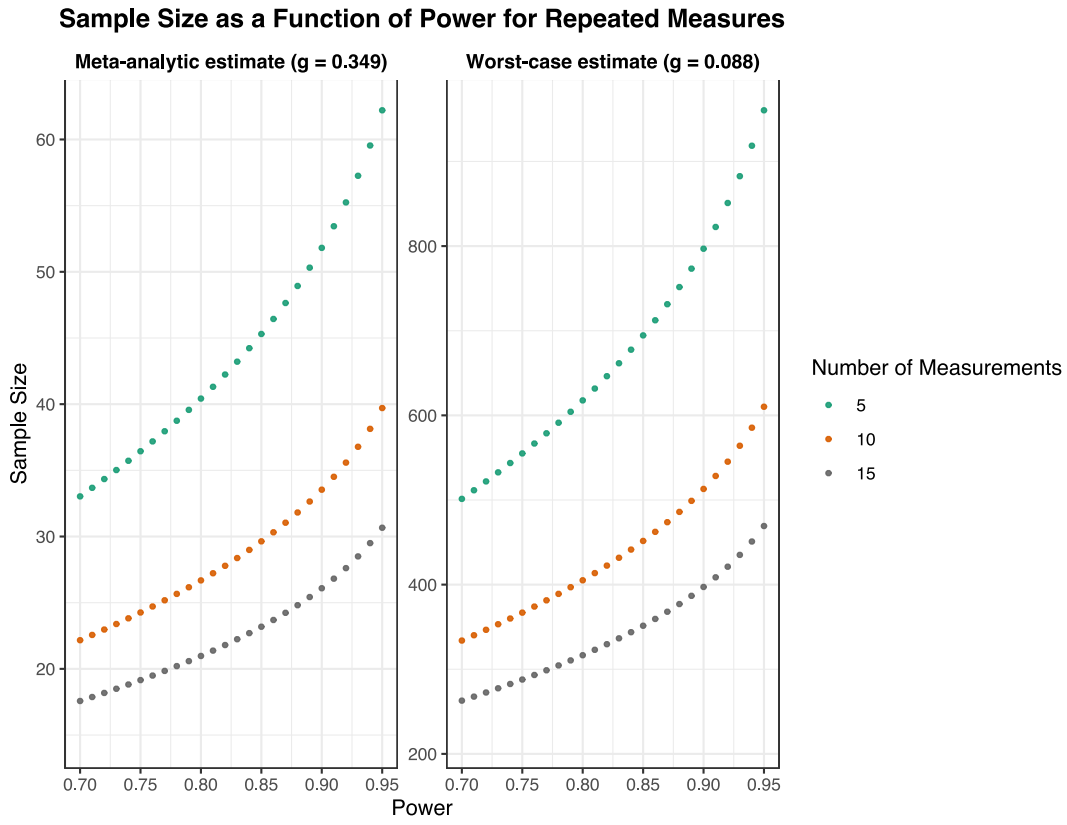


Figure 6: Plot of sample size as a function of power assuming a correlation among within-subject measurements of 0.5. Note the different scales for each of the y-axes. The data for this theoretical plot were made using G* Power (Faul, Erdfelder, Buchner, & Lang, 2009).

4.2.3: Study Recommendations

A sample within some of the above size ranges would be impractical to pursue given the current state of lab-based infancy research; however, there are other strategies to increase power and reduce the required sample size. One strategy involves improving measures of infant preference, such as by adding neuroimaging, eye tracking, or electrophysiological measures (e.g. Bristow et al., 2008; Grossmann, Missana, Friederici, & Ghazanfar, 2012; Shaw, Baart, Depowski, & Bortfeld, 2015), which are more robust to factors affecting infant attention and thereby increase the power of studies. Another strategy could involve a large-scale, cross-lab collaborative approach (e.g. ManyBabies Consortium, 2020), which would offer the additional benefit of enabling quantification of between-lab sources of heterogeneity in the investigation of this aspect of infant cognition. A more tenable strategy involves engaging in open science practices. In addition to allowing for multi-lab replication attempts and reanalysis of data, the availability of materials, data and scripts from other studies can, over time, contribute to a more comprehensive estimate of the role of experimental methods in infancy research. In line with these recommendations for data sharing, we encourage researchers with unpublished and published work to contribute their experimental results to the open MetaLab repository, as the meta-analytic studies reported here are restricted to those that were discoverable through our search criteria.

4.2.4: Open Science & Preregistration

Infancy research, moreover, exhibits particular susceptibility to publication bias due to the prevalence of small sample sizes and noisy measurements that occur as a product of the unique recruitment and testing challenges associated with infants (Bergmann et al., 2018). The above finding showing the meta-analytic effect under investigation to be moderately sensitive to publication bias advocates for researchers to engage in preregistration of future studies on infants' ability to perceive audio-visual congruence. By encouraging researchers to make a priori decisions on the rationale, design and analysis for the study, preregistration can work against publication bias, counteract questionable research practices, increase the proportion of studies that provide informative results, and improve transparency and openness in the field (Havron, Bergmann & Tsuji, 2020). Although we recommend preregistration as a tool to think more deeply about study design and analytic choices, this does not entail that motivated deviations and explicitly exploratory analyses should be discouraged (cf. Szollosi et al., 2019; Devezer et al., 2020; Navarro, 2019).

5.0: Conclusion

This meta-analysis provides moderate evidence that infants can detect audio-visual congruence for speech at an early point in the developmental process. While good progress has been made in investigating a wide variety of relevant aspects of infants' ability to combine speech information across modalities, as indicated by the citation network analyses, this meta-analysis highlights a number of relevant issues. In particular, there is a need i) to engage in open science practices and preregistration as well as to use more robust measures in order to increase the replicability of results, and ii) to conduct further studies exploring the effects of infants' individual patterns of vocal production as well as how infants adapt their allocation of attention in response to different degrees of task difficulty and stimulus synchronicity. We hope this paper advances the study of infants' ability to perceive multimodal congruence by promoting transparent, cumulative research that can contribute to stronger theories of infant language development.

6.0: Acknowledgements

We would like to acknowledge the comments, feedback and suggestions from the editor and anonymous reviewers who all substantially improved the paper. We would also like to acknowledge and thank David Lewkowicz for his helpful comments and useful feedback on the first versions of this paper. The authors declare no conflicts of interest with regard to the funding source for this study.

7.0: List of References

Aldridge, M. A., Braga, E. S., Walton, G. E., & Bower, T. (1999). The intermodal representation of speech in newborns. *Developmental Science, 2*(1), 42-46.

Altvater-Mackensen, N., Mani, N., & Grossmann, T. (2016). Audiovisual speech perception in infancy: The influence of vowel identity and infants' productive abilities on sensitivity to (mis)matches between auditory and visual speech cues. *Dev Psychol, 52*(2), 191-204. doi:10.1037/a0039964

Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. *Journal of informetrics, 11*(4), 959-975.

Azur, M. J., Stuart, E. A., Frangakis, C., & Leaf, P. J. (2011). Multiple imputation by chained equations: what is it and how does it work? *International journal of methods in psychiatric research, 20*(1), 40-49.

Bahrick, L. E., Hernandez-Reif, M., & Flom, R. (2005). The development of infant learning about specific face-voice relations. *Developmental psychology, 41*(3), 541-552. doi:10.1037/0012-1649.41.3.541

Bahrick, L. E., Netto, D., & Hernandez-Reif, M. (1998). Intermodal perception of adult and child faces and voices by infants. *Child development, 69*(5), 1263-1275. doi:10.2307/1132264

Bahrick, L. E., & Lickliter, R. (2000). Intersensory redundancy guides attentional selectivity and perceptual learning in infancy. *Developmental psychology*, *36*(2), 190-201.

Bahrick, L. E., & Lickliter, R. (2012). The role of intersensory redundancy in early perceptual, cognitive, and social development. *Multisensory development*, 183-206.

Berger, S. E. (2004). Demands on finite cognitive capacity cause infants' perseverative errors. *Infancy*, *5*(2), 217-238.

Bergmann, C., Tsuji, S., Piccinini, P. E., Lewis, M. L., Braginsky, M., Frank, M. C., & Cristia, A. (2018). Promoting replicability in developmental research through meta-analyses: Insights from language acquisition research. *Child Development, 89*(6), 1996–2009. https://doi.org/10.1111/cdev.13079

Bremner, A. J., Lewkowicz, D. J., & Spence, C. (Eds.). (2012). *Multisensory development*. Oxford University Press.

Bristow, D., Dehaene-Lambertz, G., Mattout, J., Soares, C., Gliga, T., Baillet, S., & Mangin, J. F. (2009). Hearing Faces: How the Infant Brain Matches the Face It Sees with the Speech It Hears. *Journal of Cognitive Neuroscience, 21*(5), 905-921. doi:10.1162/jocn.2009.21076

Bürkner, P.-C. (2017). brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software, 80*(1), 1-28.

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

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software, 76*(1).

Champely, S. (2016). pwr: Basic functions for power analysis. R package version 1.2-0. In.

Coulon, M., Hemimou, C., & Streri, A. (2013). Effects of seeing and hearing vowels on neonatal facial imitation. *Infancy, 18*(5), 782-796.

Desjardins, R. N., & Werker, J. F. (2004). Is the integration of heard and seen speech mandatory for infants?. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, *45*(4), 187-203.

DePaolis, R. A., Vihman, M. M., & Keren-Portnoy, T. (2011). Do production patterns influence the processing of speech in prelinguistic infants? *Infant Behavior and Development, 34*(4), 590-601.

DePaolis, R. A., Vihman, M. M., & Nakai, S. (2013). The influence of babbling patterns on the processing of speech. *Infant Behav Dev, 36*(4), 642-649. doi:10.1016/j.infbeh.2013.06.007

Devezer, B., Navarro, D. J., Vandekerckhove, J., & Buzbas, E. O. (2020). The case for formal methodology in scientific reform. BioRxiv. https://doi.org/10.1101/2020.04.26.048306

Dick, A. S., Solodkin, A., & Small, S. L. (2010). Neural development of networks for audiovisual speech comprehension. *Brain and language*, *114*(2), 101-114.

Dorn, K., Weinert, S., & Falck-Ytter, T. (2018). Watch and listen - A cross-cultural study of audio-visual-matching behavior in 4.5-month-old infants in German and Swedish talking faces. *Infant Behavior & Development, 52*, 121-129. doi:10.1016/j.infbeh.2018.05.003

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior research methods*, *41*(4), 1149-1160.

Fernández-Castilla, B., Jamshidi, L., Declercq, L., Beretvas, S. N., Onghena, P., & Van den Noortgate, W. (2020). The application of meta-analytic (multi-level) models with multiple random effects: A systematic review. *Behavior Research Methods*, 1-22.

Féron, J., Gentaz, E., & Streri, A. (2006). Evidence of amodal representation of small numbers across visuo-tactile modalities in 5-month-old infants. *Cognitive Development, 21*(2), 81-92.

Filippetti, M. L., Johnson, M. H., Lloyd-Fox, S., Dragovic, D., & Farroni, T. (2013). Body perception in newborns. *Current Biology, 23*(23), 2413-2416.

Gelman, A., Simpson, D., & Betancourt, M. (2017). The prior can often only be understood in the context of the likelihood. *Entropy*, *19*(10), 555.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C.C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.C. and Modrák, M. (2020). Bayesian workflow. *arXiv preprint arXiv:2011.01808*.

Gibson, E.J. (1969). *Principles of perceptual learning and development.* Appleton-Century-Crofts.

Groothuis-Oudshoorn, K., & Van Buuren, S. (2011). Mice: multivariate imputation by chained equations in R. *J Stat Softw, 45*(3), 1-67.

Grossmann, T., Missana, M., Friederici, A. D., & Ghazanfar, A. A. (2012). Neural correlates of perceptual narrowing in cross-species face-voice matching. *Dev Sci, 15*(6), 830-839. doi:10.1111/j.1467-7687.2012.01179.x

Guellai, B., Streri, A., Chopin, A., Rider, D., & Kitamura, C. (2016). Newborns' Sensitivity to the Visual Aspects of Infant-Directed Speech: Evidence From Point-Line Displays of Talking Faces. *Journal of Experimental Psychology-Human Perception and Performance, 42*(9), 1275-1281. doi:10.1037/xhp0000208

Guo, Y., Logan, H. L., Glueck, D. H., & Muller, K. E. (2013). Selecting a sample size for studies with repeated measures. *BMC medical research methodology*, *13*(1), 100. doi: https://doi.org/10.1186/1471-2288-13-100

Havron, N., Bergmann, C., & Tsuji, S. (2020). Preregistration in infant research—A primer. *Infancy*, *25*(5), 734-754.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological methods, 3*(4), 486-504.

Hickok, G., Houde, J., & Rong, F. (2011). Sensorimotor integration in speech processing: computational basis and neural organization. *Neuron*, *69*(3), 407-422.

Hillairet de Boisferon, A. H., Dupierrix, E., Quinn, P. C., Lœvenbruck, H., Lewkowicz, D. J., Lee, K., & Pascalis, O. (2015). Perception of Multisensory Gender Coherence in 6- and 9-month-old Infants. *Infancy, 20*(6), 661-674. doi:10.1111/infa.12088

Hillairet de Boisferon, A., Tift, A. H., Minar, N. J., & Lewkowicz, D. J. (2017). Selective attention to a talker's mouth in infancy: role of audiovisual temporal synchrony and linguistic experience. *Developmental science*, *20*(3), e12381.

Hunter, M. A., & Ames, E. W. (1988). A multifactor model of infant preferences for novel and familiar stimuli. *Advances in Infancy Research, 5*, 69–95.

Imada, T., Zhang, Y., Cheour, M., Taulu, S., Ahonen, A., & Kuhl, P. K. (2006). Infant speech perception activates Broca's area: a developmental magnetoencephalography study. *Neuroreport*, *17*(10), 957-962.

Jylänki, P., Vanhatalo, J., & Vehtari, A. (2011). Robust Gaussian Process Regression with a Student-t Likelihood. *Journal of Machine Learning Research, 12*(11), 3227-3257.

Kidd, C., Piantadosi, S. T., & Aslin, R. N. (2012). The Goldilocks effect: Human infants allocate attention to visual sequences that are neither too simple nor too complex. *PloS one, 7*(5), e36399.

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan*. Academic Press.

Kubicek, C., Hillairet de Boisferon, A. H., Dupierrix, E., Loevenbruck, H., Gervain, J., & Schwarzer, G. (2013). Face-scanning behavior to silently-talking faces in 12-month-old infants: The impact of pre-exposed auditory speech. *International Journal of Behavioral Development, 37*(2), 106-110.

Kubicek, C., Hillairet de Boisferon, A. H., Dupierrix, E., Pascalis, O., Loevenbruck, H., Gervain, J., & Schwarzer, G. (2014). Cross-Modal Matching of Audio-Visual German and French Fluent Speech in Infancy. *PloS one, 9*(2). doi:10.1371/journal.pone.0089275

Kubicek, C., Gervain, J., Hillairet de Boisferon, A., Pascalis, O., Lœvenbruck, H., & Schwarzer, G. (2014). The influence of infant-directed speech on 12-month-olds' intersensory perception of fluent speech. *Infant Behav Dev, 37*(4), 644-651. doi:10.1016/j.infbeh.2014.08.010

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

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

Kuhl, P. K., & Meltzoff, A. N. (1996). Infant vocalizations in response to speech: vocal imitation and developmental change. *J Acoust Soc Am, 100*(4 Pt 1), 2425-2438. doi:10.1121/1.417951

Kuhl, P. K., Ramírez, R. R., Bosseler, A., Lin, J. F. L., & Imada, T. (2014). Infants' brain responses to speech suggest analysis by synthesis. *Proceedings of the National Academy of Sciences*, *111*(31), 11238-11245.

Kuhl, P. K., Stevens, E., Hayashi, A., Deguchi, T., Kiritani, S., & Iverson, P. (2006). Infants show a facilitation effect for native language phonetic perception between 6 and 12 months. *Developmental science*, *9*(2), F13-F21.

Kushnerenko, E., Teinonen, T., Volein, A., & Csibra, G. (2008). Electrophysiological evidence of illusory audiovisual speech percept in human infants. *Proceedings of the National Academy of Sciences*, *105*(32), 11442-11445.

Legerstee, M. (1990). Infants use multimodal information to imitate speech sounds. *Infant Behavior and Development, 13*(3), 343-354.

Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, *128*(7), 912-928.

Lewkowicz, D. J. (2000). The development of intersensory temporal perception: an epigenetic systems/limitations view. *Psychological bulletin*, *126*(2), 281-308.

Lewkowicz, D. J. (2010). Infant perception of audio-visual speech synchrony. *Developmental psychology*, *46*(1), 66-77.

Lewkowicz, D. J. (2014). Early experience and multisensory perceptual narrowing. *Developmental psychobiology*, *56*(2), 292-315.

Lewkowicz, D. J., & Ghazanfar, A. A. (2006). The decline of cross-species intersensory perception in human infants. *Proceedings of the National Academy of Sciences*, *103*(17), 6771-6774.

Lewkowicz, D. J., & Hansen-Tift, A. M. (2012). Infants deploy selective attention to the mouth of a talking face when learning speech. *Proceedings of the National Academy of Sciences*, *109*(5), 1431-1436.

Lewkowicz, D. J., Minar, N. J., Tift, A. H., & Brandon, M. (2015). Perception of the multisensory coherence of fluent audiovisual speech in infancy: Its emergence and the role of experience. *Journal of Experimental Child Psychology, 130*, 147-162.

Lewkowicz, D. J., & Pons, F. (2013). Recognition of amodal language identity emerges in infancy. *International Journal of Behavioral Development, 37*(2), 90-94.

Matchin, W., Groulx, K., & Hickok, G. (2014). Audiovisual speech integration does not rely on the motor system: evidence from articulatory suppression, the McGurk effect, and fMRI. *Journal of Cognitive Neuroscience*, *26*(3), 606-620.

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

Majorano, M., Vihman, M. M., & DePaolis, R. A. (2014). The relationship between infants' production experience and their processing of speech. *Language Learning and Development, 10*(2), 179-204.

ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science, 3*(1), 24-52.

Mathur, M. B., & VanderWeele, T. J. (2020). Sensitivity analysis for publication bias in meta-analyses. *Journal of the Royal Statistical Society. Series C, Applied Statistics, 69*(5), 1091-1119.

Maurer, D., & Werker, J. F. (2014). Perceptual narrowing during infancy: A comparison of language and faces. *Developmental Psychobiology*, *56*(2), 154-178.

McElreath, R. (2020). *Statistical rethinking: A Bayesian course with examples in R and Stan*. CRC press.

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

Meltzoff, A. N., & Kuhl, P. K. (1994). *Faces and speech: Intermodal processing of biologically relevant signals in infants and adults.* In D. J. Lewkowicz & R. Lickliter (Eds.), *The development of intersensory perception: Comparative perspectives* (p. 335–369). Lawrence Erlbaum Associates, Inc.

Morris, S. B. (2000). Distribution of the standardized mean change effect size for meta-analysis on repeated measures. *British Journal of Mathematical and Statistical Psychology*, *53*(1), 17-29.

Navarro, D. J. (2019). Between the devil and the deep blue sea: Tensions between scientific judgement and statistical model selection. *Computational Brain & Behavior*, *2*(1), 28-34.

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

Patterson, M. L., & Werker, J. F. (2002). Infants' ability to match dynamic phonetic and gender, information in the face and voice. *Journal of Experimental Child Psychology, 81*(1), 93-115. doi:10.1006/jecp.2001.2644

Pejovic, J., Yee, E., & Molnar, M. (2020). Speaker matters: Natural inter-speaker variation affects 4-month-olds' perception of audio-visual speech. *First Language, 40*(2), 113-127. doi:10.1177/0142723719876382

Pickens, J., Field, T., Nawrocki, T., Martinez, A., Soutullo, D., & Gonzalez, J. (1994). Full-term and Preterm Infants' Perception of Face-Voice Synchrony. *Infant Behavior and Development, 17*(4), 447-455. doi:10.1016/0163-6383(94)90036-1

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

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.

Poulindubois, D., Serbin, L. A., Kenyon, B., & Derbyshire, A. (1994). Infants' Intermodal Knowledge about Gender. *Developmental psychology, 30*(3), 436-442. doi:10.1037/0012-1649.30.3.436

R Core Team. (2013). R: A Language Environment for Statistical Computing: Vienna, Austria: R Foundation for Statistical Computing.

RStudio Team (2018). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA

Richoz, A. R., Quinn, P. C., Hillairet de Boisferon, A., Berger, C., Loevenbruck, H., Lewkowicz, D. J., Pascalis, O. (2017). Audio-Visual Perception of Gender by Infants Emerges Earlier for Adult-Directed Speech. *PloS one, 12*(1), e0169325. doi:10.1371/journal.pone.0169325

Rosenblum, L. D., Schmuckler, M. A., & Johnson, J. A. (1997). The McGurk effect in infants. *Perception & psychophysics*, *59*(3), 347-357.

Rose, S. A., Gottfried, A. W., Melloy-Carminar, P., & Bridger, W. H. (1982). Familiarity and novelty preferences in infant recognition memory: Implications for information processing. *Developmental Psychology*, *18*(5), 704-713.

Sann, C., & Streri, A. (2007). Perception of object shape and texture in human newborns: evidence from cross-modal transfer tasks. *Developmental Science, 10*(3), 399-410.

Segal, O., Hejli-Assi, S., & Kishon-Rabin, L. (2016). The effect of listening experience on the discrimination of/ba/and/pa/in Hebrew-learning and Arabic-learning infants. *Infant Behavior and Development*, *42*, pp. 86-99.

Shaw, K., Baart, M., Depowski, N., & Bortfeld, H. (2015). Infants' Preference for Native Audiovisual Speech Dissociated from Congruency Preference. *PloS one, 10*(4). doi:10.1371/journal.pone.0126059

Sterne, J. A., White, I. R., Carlin, J. B., Spratt, M., Royston, P., Kenward, M. G., . . . Carpenter, J. R. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj, 338*:b2393.

Stewart, L. A., Clarke, M., Rovers, M., Riley, R. D., Simmonds, M., Stewart, G., & Tierney, J. F. (2015). Preferred reporting items for a systematic review and meta-analysis of individual participant data: the PRISMA-IPD statement. *Jama, 313*(16), 1657-1665.

Streri, A., Coulon, M., Marie, J., & Yeung, H. H. (2016). Developmental Change in Infants' Detection of Visual Faces that Match Auditory Vowels. *Infancy, 21*(2), 177-198. doi:10.1111/infa.12104

Szollosi, A., Kellen, D., Navarro, D. J., Shiffrin, R., van Rooij, I., Van Zandt, T., & Donkin, C. (2019). Is preregistration worthwhile. *Trends in Cognitive Sciences*, *24*(2), 94-95.

Trehub, S. E., Plantinga, J., & Brcic, J. (2009). Infants detect cross-modal cues to identity in speech and singing. *Ann N Y Acad Sci, 1169*, 508-511. doi:10.1111/j.1749-6632.2009.04851.x

Tsuji, S., Cristia, A., Frank, M. C., & Bergmann, C. (2019). Addressing publication bias in meta-analysis: Empirical findings from community-augmented meta-analyses of infant language development. https://doi.org/10.31222/osf.io/q5axy

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing, 27*(5), 1413-1432.

Von Holzen, K., & Bergmann, C. (2018). A Meta-Analysis of Infants' Mispronunciation Sensitivity Development. *Annual Conference of the Cognitive Science Society. Cognitive Science Society (U.S.). Conference*, *2018*, 1157–1162.

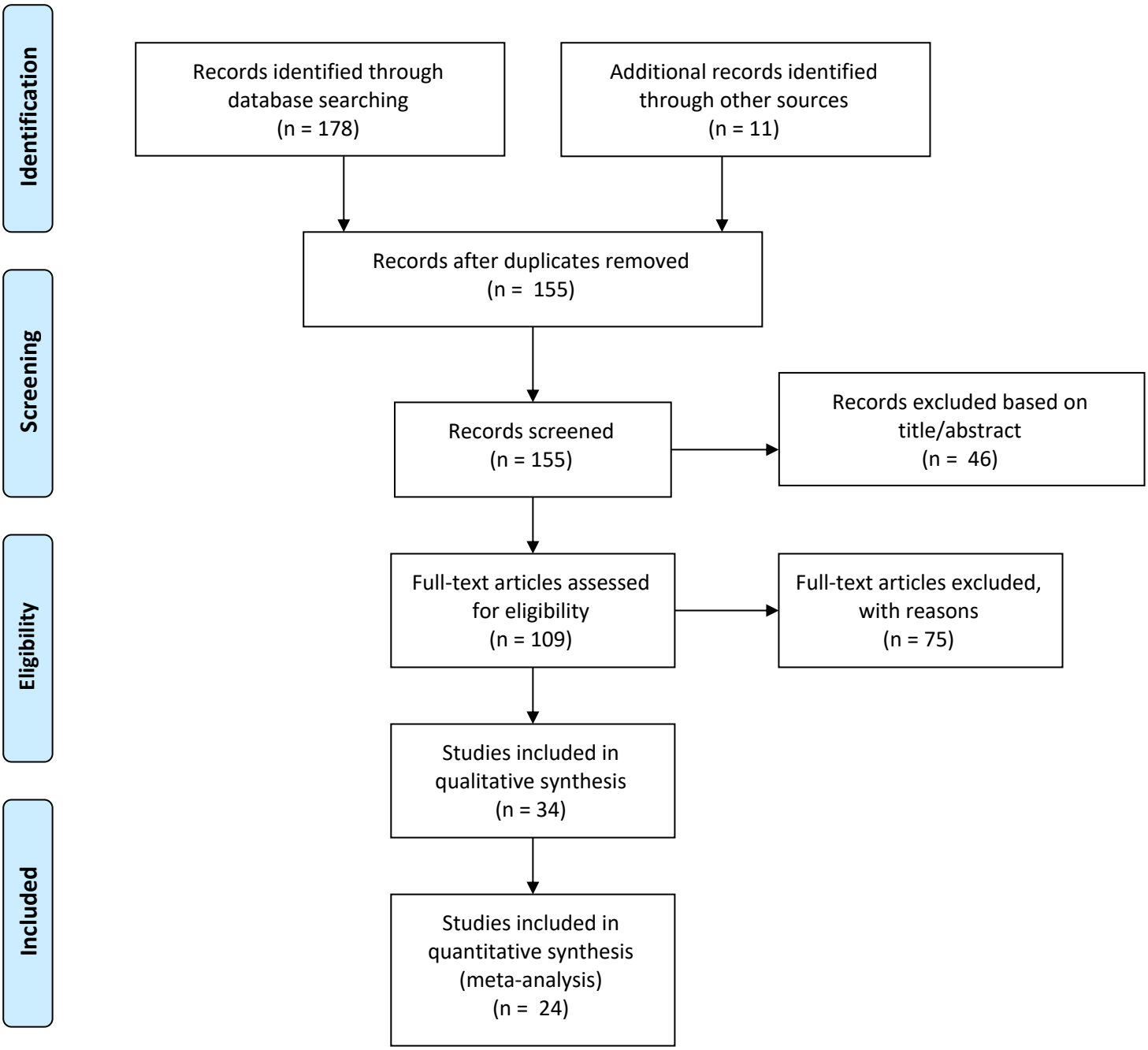
Walker-Andrews, A. S., Bahrick, L. E., Raglioni, S. S., & Diaz, I. (1991). Infants' bimodal perception of gender. *Ecological Psychology, 3*(2), 55-75.

Werker JF, Gervain J (2013) Speech Perception in Infancy: A foundation for Language Acquisition. In P. D. Zelazo (Ed.), *The Oxford Handbook of Developmental Psychology* (pp. 909-925). Oxford University Press.

Yeung, H. H., & Werker, J. F. (2013). Lip movements affect infants' audiovisual speech perception. *Psychol Sci, 24*(5), 603-612. doi:10.1177/0956797612458802

Appendix A

PRISMA 2009 Flow Diagram



Appendix B

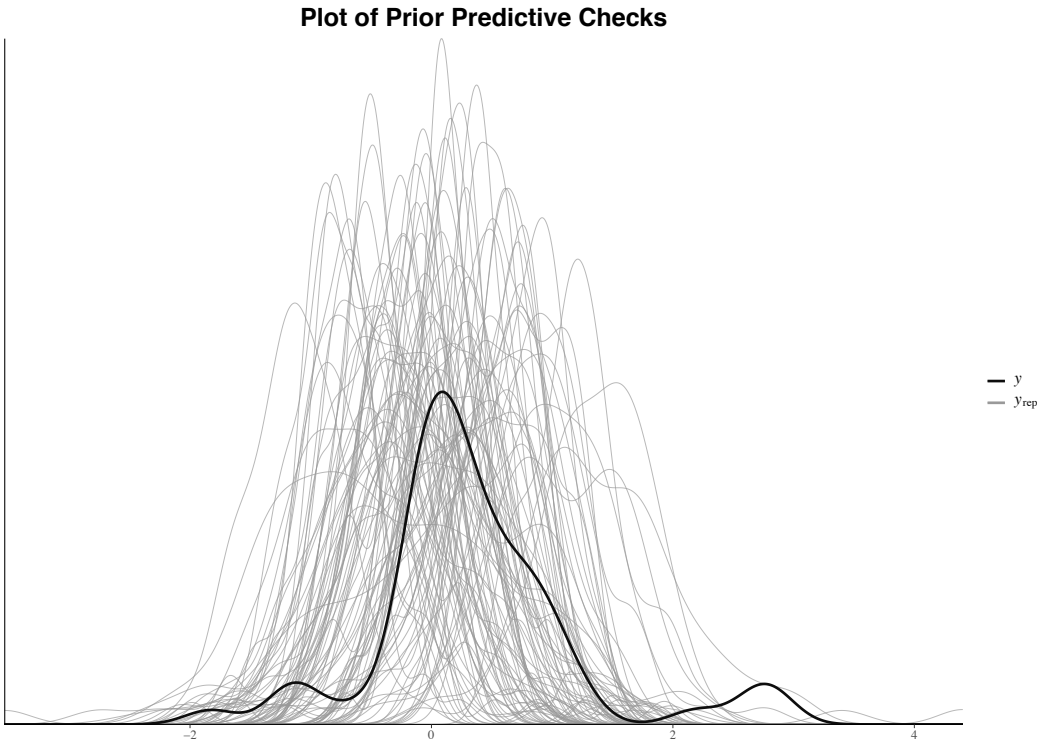


Figure 7: Plot of the prior predictive checks (grey) and observed meta-analytic data (black).

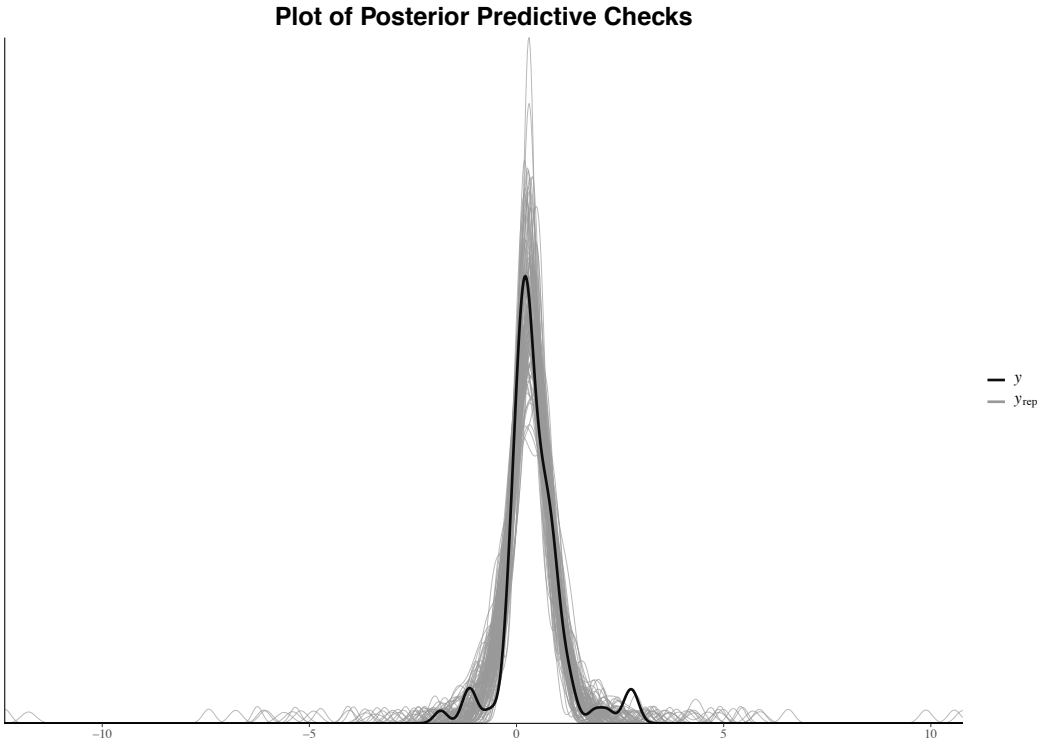


Figure 8: Plot of the posterior predictive checks (grey) and observed meta-analytic data (black).

Appendix C

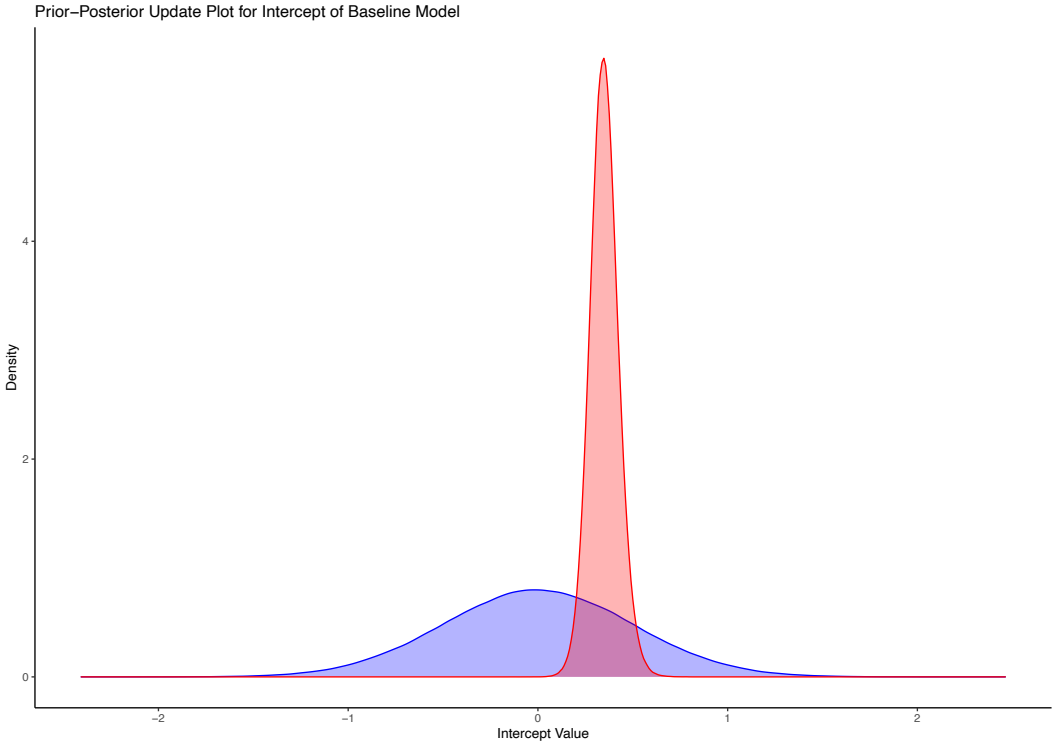


Figure 9: Plot of the prior (blue) and posterior (red) distributions for the intercept in the baseline model.

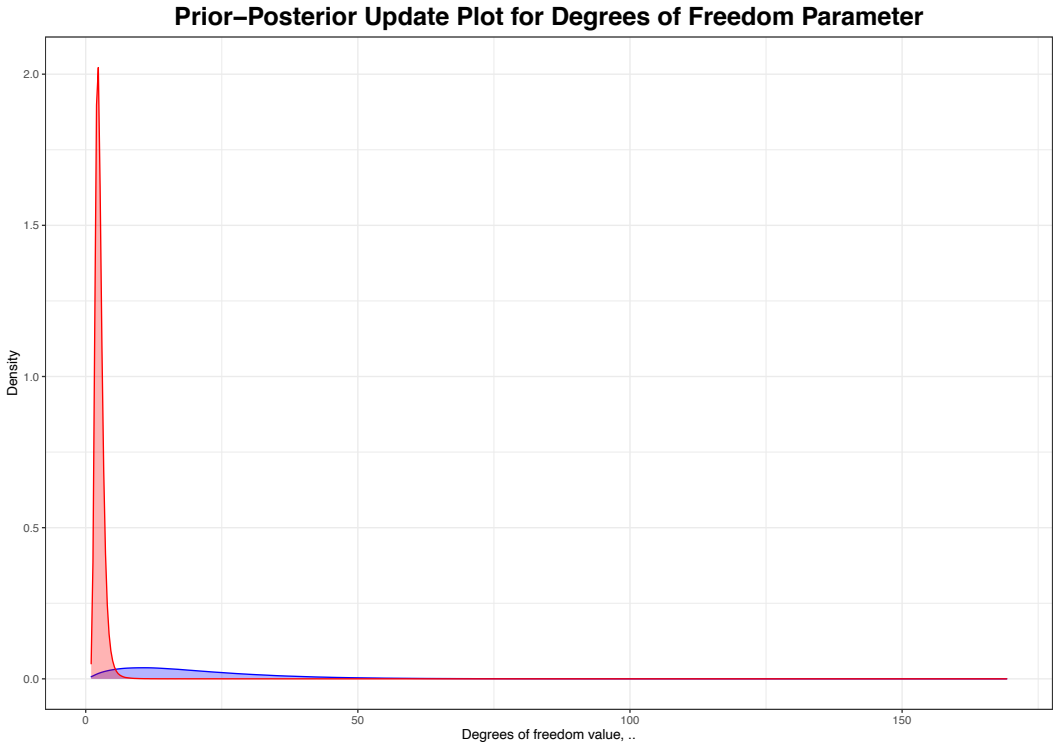


Figure 10: Plot of the prior (blue) and posterior (red) distributions for the degrees of freedom parameter, ν, of the Student's t-distribution in the baseline model.

Figure 11: Plot of the prior (blue) and posterior (red) distributions for the within-study standard deviation in the baseline model.

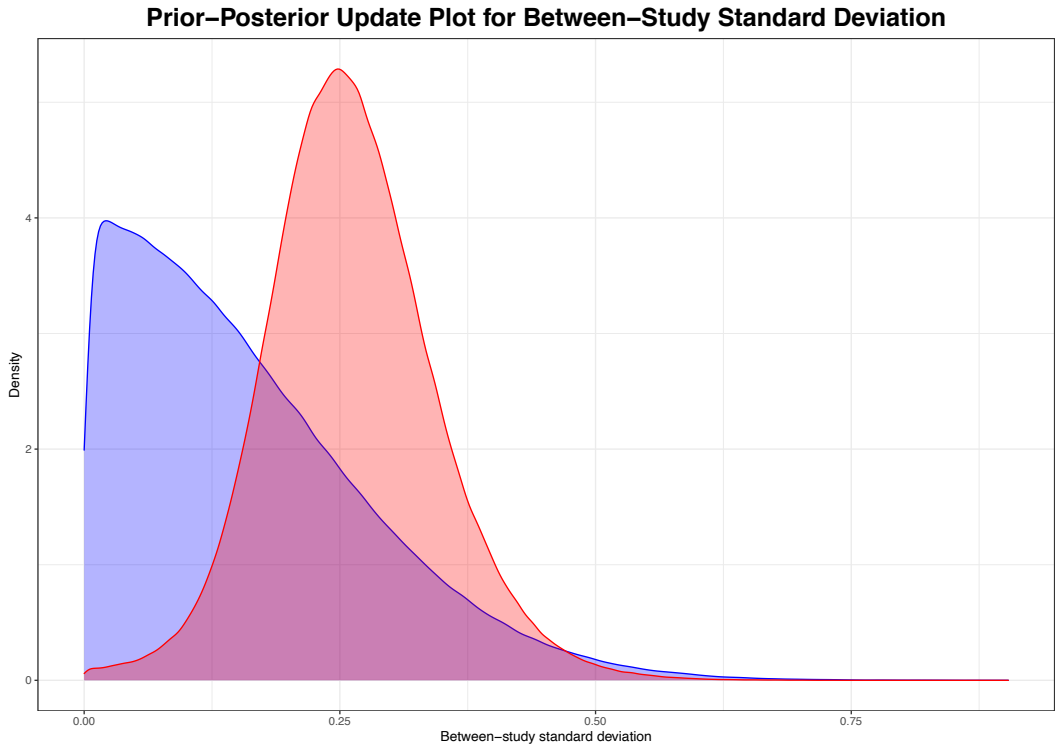


Figure 12: Plot of the prior (blue) and posterior (red) distributions for the between-study standard deviation in the baseline model.

Appendix D

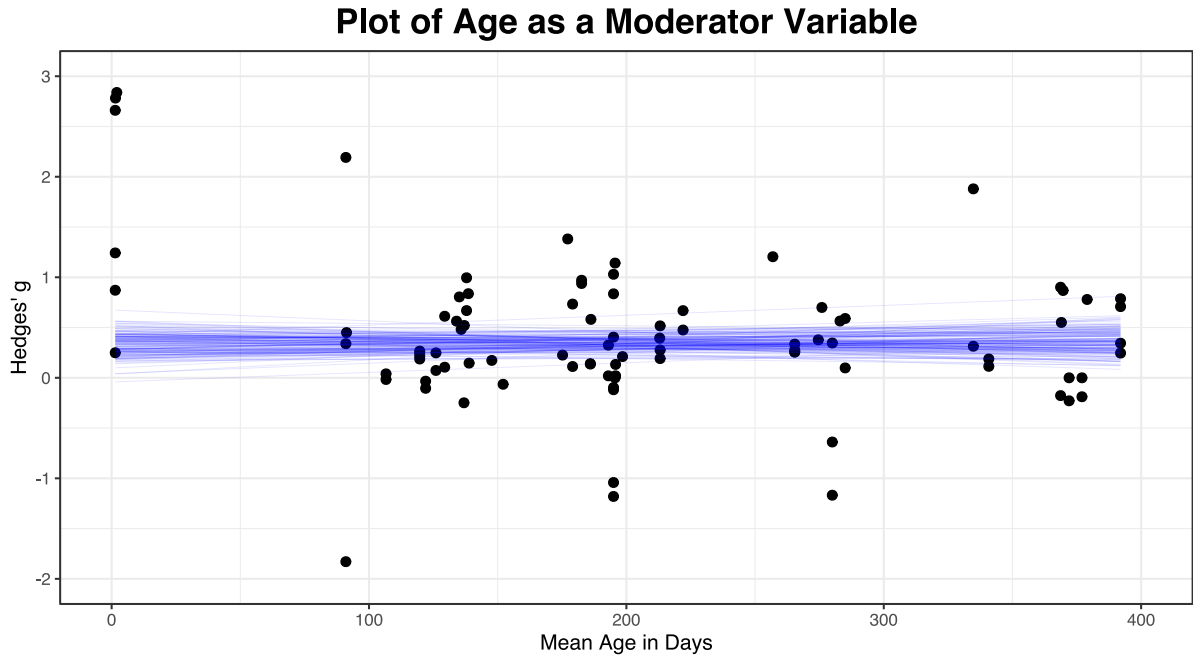


Figure 13: Spaghetti plot showing 150 posterior model predictions for the moderator variable of age. The data indicate no clear developmental patterns in neither a positive nor a negative direction.

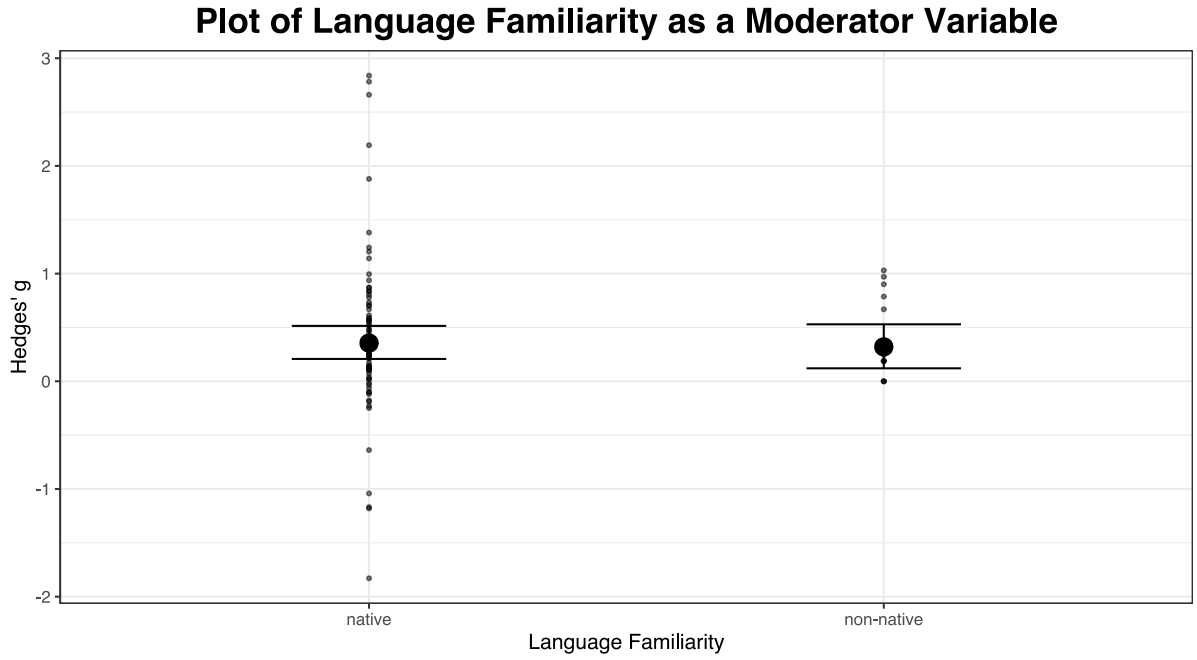


Figure 14: Plot showing the moderator variable of language familiarity. Although there are fewer tests of non-native stimuli, they pattern close to those of the native stimuli.

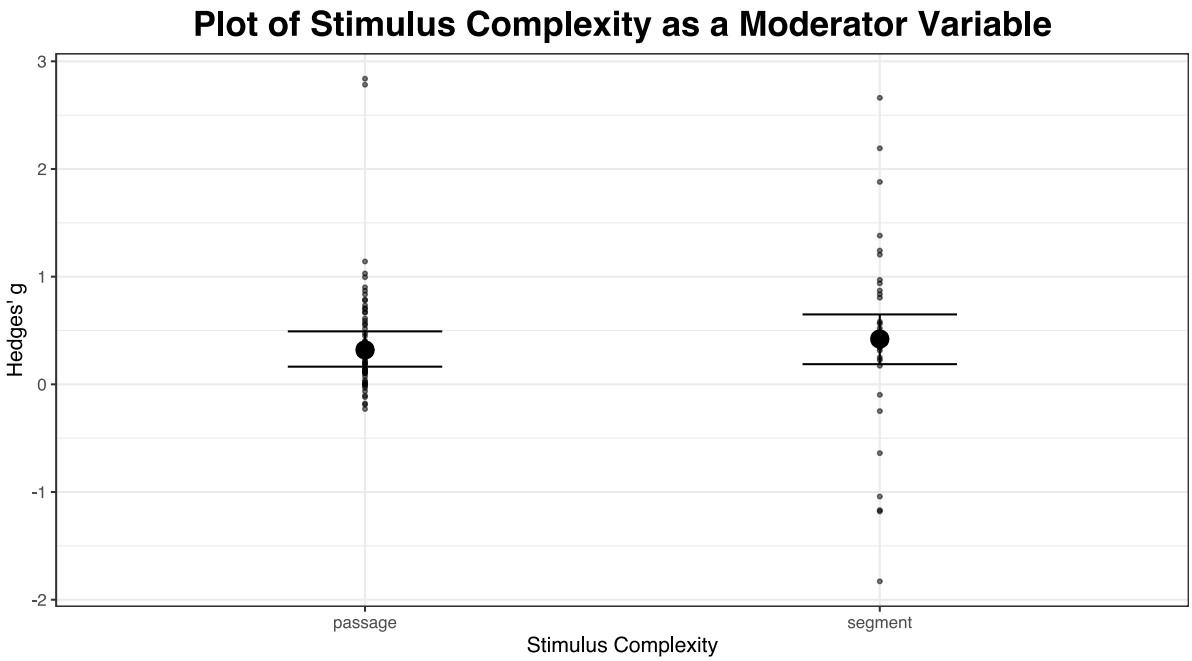


Figure 15: Plot showing the moderator variable of stimulus complexity.

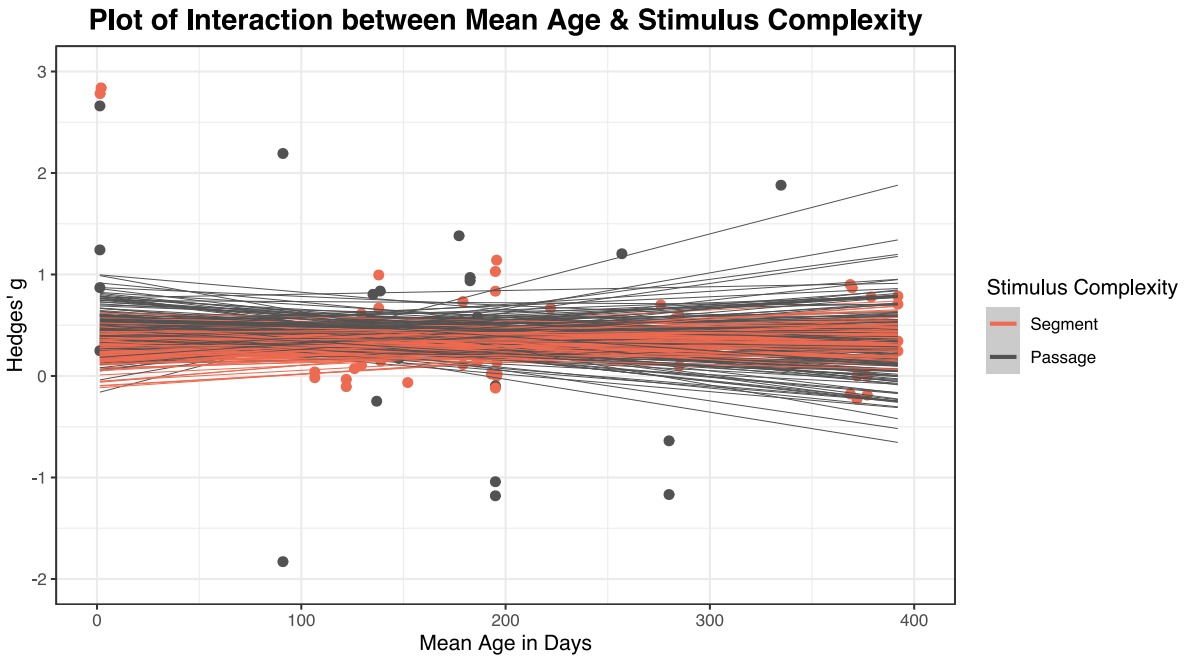


Figure 16: Spaghetti plot showing 150 posterior model predictions for the interaction between age and stimulus complexity. The data indicate no clear developmental patterns in neither a positive nor a negative direction.

Appendix E

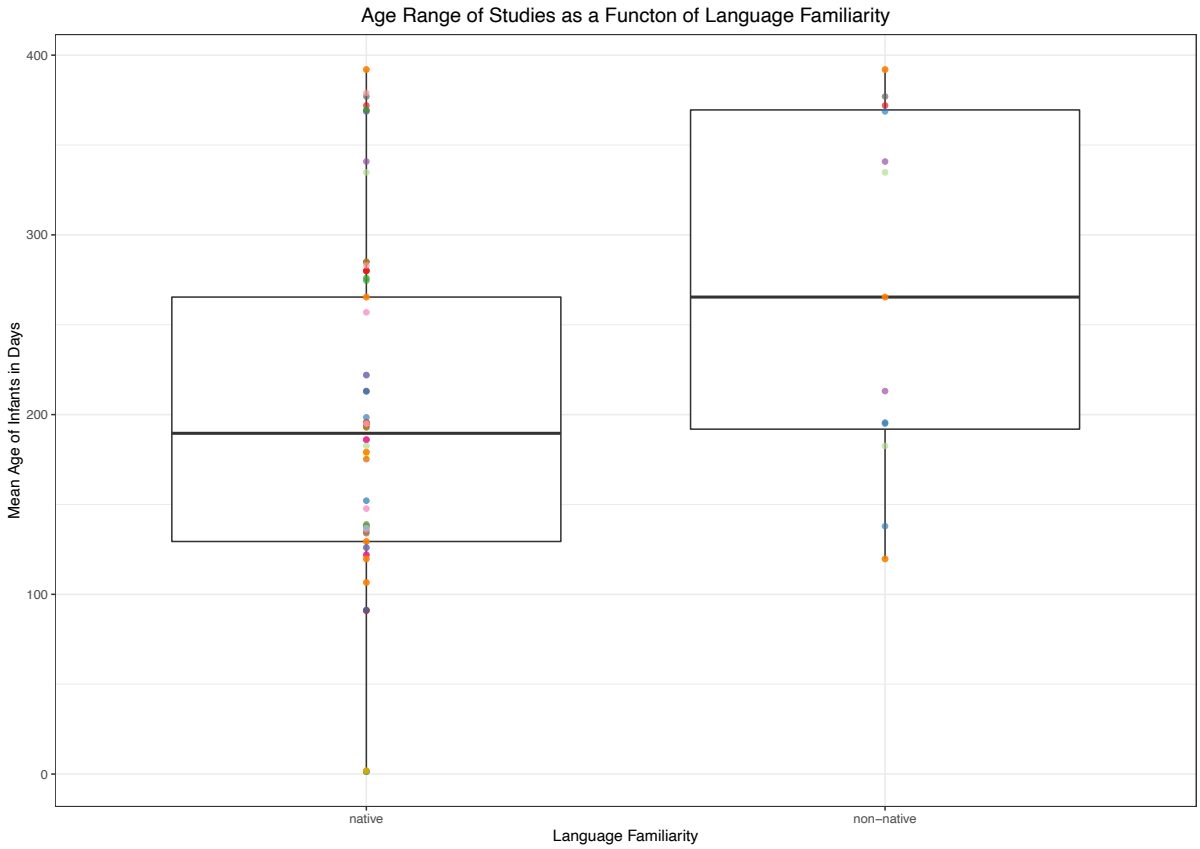


Figure 17: Boxplot of the age ranges of experimental studies investigating infants' ability to perceive audio-visual congruence for native and non-native speech. The points represent individual experiments and the colours of the points indicate individual studies (N=24).

Appendix F

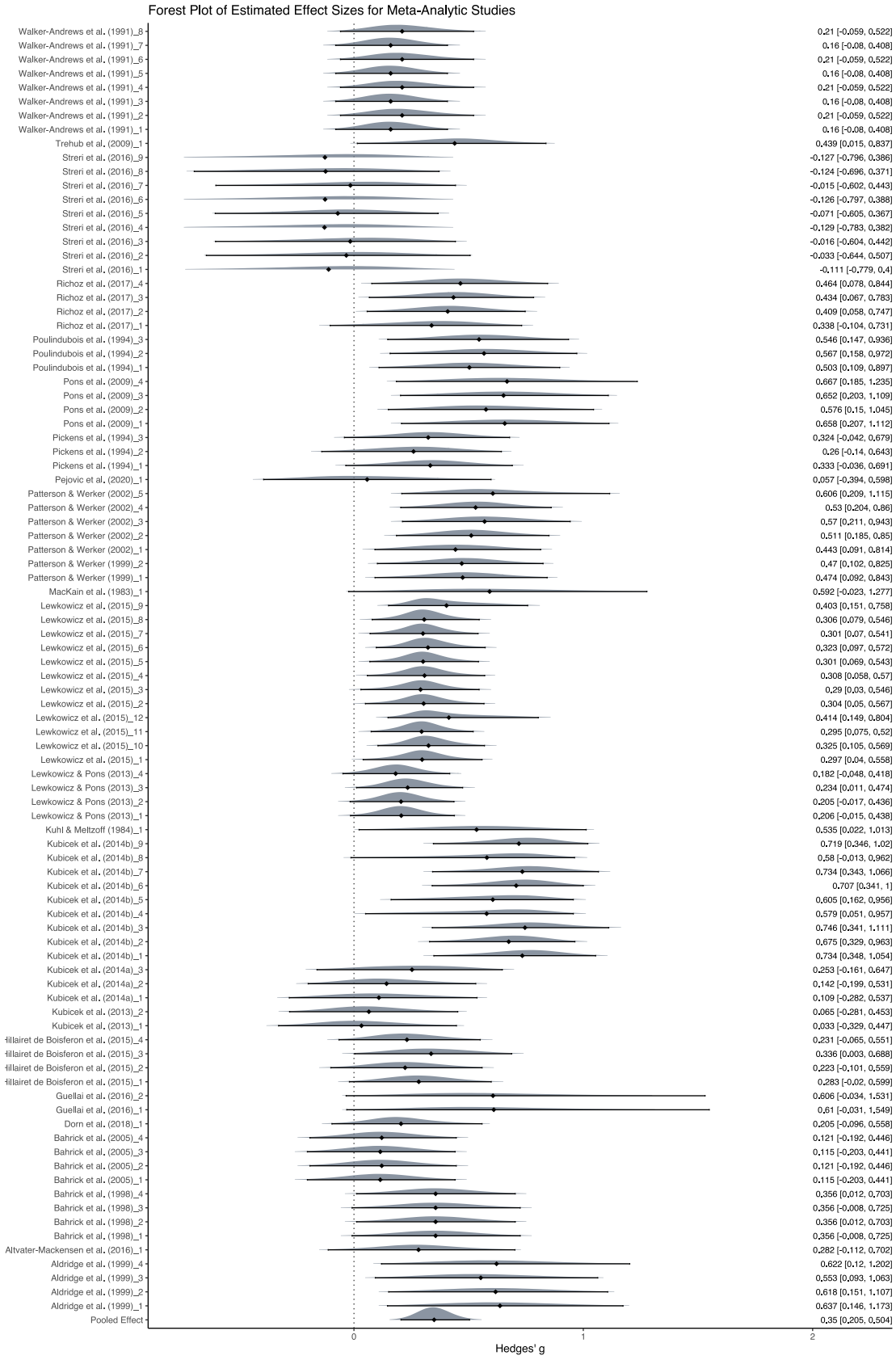


Figure 18: Forest plot of estimated effect sizes for infants' detection of audio-visual congruence for each experiment.