# The development of infants' responses to mispronunciations: A Meta-Analysis

Katie Von Holzen[1,2,3] & Christina Bergmann[4,5]

[1] Lehrstuhl Linguistik des Deutschen, Schwerpunkt Deutsch als Fremdsprache/Deutsch als Zweitsprache, Technische Universität Dortmund

[2] Department of Hearing and Speech Sciences, University of Maryland, USA

[3] Laboratoire Psychologie de la Perception, Université Paris Descartes

[4] Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands

[5] LSCP, Departement d'Etudes Cognitives, ENS, EHESS, CNRS, PSL Research University

## Author Note

The authors each declare that they have no conflict of interest.

Correspondence concerning this article should be addressed to Katie Von Holzen, Emil-Figge-Straße 50, 44221 Dortmund, Germany. E-mail: katie.m.vonholzen@gmail.com

## Abstract

As they develop into mature speakers of their native language, infants must not only learn words but also the sounds that make up those words. To do so, they must strike a balance between accepting the speaker's variation (e.g. mood, voice, accent), but appropriately rejecting variation when it changes a word's meaning (e.g. cat vs. hat). This meta-analysis focuses on studies investigating infants' ability to detect mispronunciations in familiar words, or mispronunciation sensitivity. Our goal was to evaluate the development of mispronunciation sensitivity in infancy as well as explore the role of experimental manipulations related to theoretical questions and analysis choices. The results show that although infants are sensitive to mispronunciations, they still accept these altered forms as labels for target objects. Interestingly, this ability is not modulated by age or vocabulary size, suggesting that a mature understanding of native language phonology may be present in infants from an early age, possibly before the vocabulary explosion. The results also support several theoretical assumptions made in the literature, such as sensitivity to mispronunciation size and position of the mispronunciation modulate mispronunciation sensitivity. We also shed light on the impact of data analysis choices that may lead to different conclusions regarding the development of infants' mispronunciation sensitivity. Our paper concludes with recommendations for improved practice in testing infants' word and sentence processing on-line.

*Keywords:* language acquisition; mispronunciation sensitivity; word recognition; meta-analysis; lexicon; infancy

# The development of infants' responses to mispronunciations: A Meta-Analysis

## Introduction

At the turn of the millenium, infant language acquisition researchers had begun to explore the phonetic information that infants attend to while segmenting words from the speech stream (Jusczyk & Aslin, 1995) and learning minimal pairs (Stager & Werker, 1997). Swingley and Aslin (2000) expanded this exploration to infants's existing representations, investigating how infants interpret phonological variation in familiar word recognition. American-English learning 18- to 23-month-olds were presented with pairs of images (e.g. a baby and a dog) and their eye movements to each image were recorded. Infants either heard the correct label (e.g. "baby") or a mispronounced label (e.g. "vaby") for one of the images. Although infants looked at the correct, target image in response to both types of labels, correct labels elicited more looks to the correct, target image than mispronounced labels. Swingley and Aslin (2000) concluded that already before the second birthday, children's representations for familiar words are phonologically well specified.

In a mature phono-lexical system, word recognition must balance flexibility to slight variation (e.g., speaker identity, accented speech) while distinguishing between phonological contrasts that differentiate words in a given language (e.g. cat-hat). The study of Swingley and Aslin (2000) as well as subsequent studies examining infants' application of phonological categories during word recognition through the mispronunciation sensitivity paradigm probe this latter distinction. Phonological contrasts relevant for the infant language-learner are determined by their native language. For an infant learning Catalan, the vowel contrast /e/-/E/ signifies a change in meaning, whereas this is not the case for an infant learning Spanish. These contrasts are therefore not innate, but must be learned. In this meta-analysis, we focus on infants' developing ability to correctly apply the phonological distinctions for their native language during word recognition. By aggregating all publicly available evidence using a meta-analysis, we can examine developmental trends making use of data from a much larger and diverse sample of infants than is possible in most single studies (see Frank, Braginsky, Yurovsky, and Marchman (2017); ManyBabiesConsortium (n.d.); for a notable exception). Before we outline the meta-analytical approach and its advantages in detail, we first discuss the proposals this study seeks to disentangle and the data supporting each of the accounts.

Research following the seminal study by Swingley and Aslin (2000) has extended mispronunciation sensitivity to infants as young as 11 months (Bergelson & Swingley, 2017), indicating that from early stages of the developing lexicon onwards, infants can and do detect mispronunciations. Regarding the change in mispronunciation sensitivity over development, however, only about half of studies have compared more than one age group on the same mispronunciation task (see Table 1) and of those, all possible patterns of development are found. This renders conclusions regarding developmental change in mispronunciation sensitivity difficult.

Several studies have found evidence for *greater* mispronunciation sensitivity as children develop (Altvater-Mackensen, 2010; Altvater-Mackensen, Feest, & Fikkert, 2014; Feest & Fikkert, 2015; Mani & Plunkett, 2007). More precisely, the difference in target looking for correct and mispronounced trials is reported to be smaller in younger infants and grows as infants develop. Other studies have found *no difference* in mispronunciation sensitivity at different ages (Bailey & Plunkett, 2002; Swingley & Aslin, 2000; Zesiger, Lozeron, Levy, & Frauenfelder, 2012). For example, Swingley and Aslin (2000) tested infants over a wide age range of 5 months (18 to 23 months). They found that age correlated with target fixations for both correct and mispronounced labels, whereas the difference between the two (mispronunciation sensitivity) did not. At least one study has found evidence for infants to become *less* sensitive to mispronunciations as they develop, showing a decrease in sensitivity from 18 to 24 months (Mani & Plunkett, 2011).

Given the diverse evidence for developmental change, or lack thereof, the question arises as to what could be driving these differences. The first words that infants learn are often not similar sounding (e.g. mama, ball, kitty; Charles-Luce & Luce, 1995) and encoding representations for these words using fine phonological detail may not be necessary. With time, more phonologically similar words are learned, which may drive a need to the infant to represent familiar words with more fine phonological detail. A growing vocabulary also reflects increased experience or familiarity with words, which may sharpen the phonological detail of their representations (Barton, Miller, & Macken, 1980). If vocabulary growth leads to an increase in the phonological specificity of infants' word representation, we should find a relationship between vocabulary size and mispronunciation sensitivity.

Yet, the majority of studies examining a potential association between mispronunciation sensitivity and vocabulary size have concluded that there is no relationship (Bailey & Plunkett, 2002; Ballem & Plunkett, 2005; Mani, Coleman, & Plunkett, 2008; Mani & Plunkett, 2007; Swingley, 2009; Swingley & Aslin, 2000, 2002; Zesiger et al., 2012). One notable exception comes from Mani and Plunkett (2010). Here, 12-month-old infants were divided into low and high vocabulary groups. High vocabulary infants showed greater sensitivity to vowel mispronunciations than low vocabulary infants, although this was not the case for consonant mispronunciations. Taken together, there is very little evidence for a role of vocabulary size in mispronunciation sensitivity. In our current meta-analysis, we include the relationship between mispronunciation sensitivity and vocabulary size to better understand this issue.

An alternative explanation is the level of task difficulty infants are presented with in mispronunciation sensitivity experiments. Although all mispronunciation sensitivity studies are generally interested in the the phonological detail with which infants represent familiar words, many studies pose more nuanced questions, such as examining the impact of number of phonological features changed or the position of the mispronunciation. Some studies may differ in their experimental design, presenting a distractor image that overlaps with the target image in the onset phoneme or a completely novel, unfamiliar distractor image. These experimental manipulations have the potential to create experimental tasks that are more or less difficult for the infant to successfully complete. The PRIMIR Framework (Processing Rich Information from Multidimensional Interactive Representations; Curtin, Byers-Heinlein, & Werker, 2011; Curtin & Werker, 2007; Werker & Curtin, 2005) describes how infants acquire and organize the incoming speech signal into phonetic and indexical detail. The ability to access and use this detail, however, is governed by the task or developmental demands (discrimination vs. word learning) probed in a particular experiment. In a particularly demanding task, such as when the target and distractor image share the same onset phoneme, infants' ability to access the phonological detail of familiar words may be restricted. If older infants are more likely to be tested using a more demaning mispronunciation sensitivity task, this may attenuate developmental effects across studies.

These manipulations may impact task demands, but they are also theoretically interesting, focusing on issues at the intersection of phonological development and lexical processing. For specific questions where we can aggregate multiple studies, we take the opportunity to shine a meta-analytic light on what modulates infants' ability to detect mispronunciations in follow-up exploatory analyses. We outline below first which nuanced questions have been frequently asked to provide a more in-depth overview of the current literature.

The first set of questions concern how infants' sensitivity is modulated by different kinds of mispronunciations. Some experiments examine infants' sensitivity to factors that change the identity of a word on a measurable level (i.e. 1-feature, 2-features, 3-features, etc.), finding graded sensitivity to both consonant (Bernier & White, 2017; Tamasi, 2016; White & Morgan, 2008) and vowel (Mani & Plunkett, 2011) feature changes. This also has consequences for task demands, as the greater the number of features changed, or *mispronunciation size*, the easier it may be to detect a mispronunciation, whereas more similar mispronunciations may be more difficult to detect.

Consonantal changes may be more disruptive to lexical processing than vowel changes in both adults (Nazzi & Cutler, 2018) and infants (Nazzi, Poltrock, & Von Holzen, 2016). In mispronunciation sensitivity, this would translate to consonant mispronunciations impairing word recognition to a greater degree than vowel mispronunciations. Yet, the handful of studies directly comparing sensitivity to consonant and vowel mispronunciations mostly find symmetry as opposed to an asymmetry between consonants and vowels for English- (Mani & Plunkett, 2007, 2010) and Danish-learning infants (Højen et al., n.d.). One study with English-learning infants did find weak evidence for greater sensitivity to consonant compared to vowel mispronunciations (Swingley, 2016). Evidence from word-learning suggests that assymetry in sensitivity to consonant and vowel changes may develop at different times depending on the infants' native language (e.g. Floccia, Nazzi, Luche, Poltrock, & Goslin, 2014; Nazzi, Floccia, Moquet, & Butler, 2009). In the current meta-analysis, we attempt to synthesize studies examining sensitivity to the *type of mispronunciation*, whether consonant or vowel, across different ages to determine whether infants generally exhibit more sensitivity to consonant compared to vowel mispronunciations in familiar word recognition as predicted by a learned account of C-bias emergence (Floccia et al., 2014; Keidel, Jenison, Kluender, & Seidenberg, 2007; Nazzi et al., 2016) and whether this is impacted by the language family of the infants' native language (Nazzi et al., 2016).

Models of spoken word processing place more or less importance on the position of a phoneme in a word. The COHORT model (Marslen-Wilson & Zwitserlood, 1989) describes lexical access in one direction, with the importance of each phoneme decreasing as its position comes later in the word. In contrast, the TRACE model (McClelland & Elman, 1986) describes lexical access as constantly updating and reevaluating the incoming speech input in the search for the correct lexical entry, and therefore can recover from word onset and to a lesser extent medial mispronunciations. The *position of mispronunciation* in the word may therefore differentially interrupt the infant's word recognition process, with onset mispronunciations resulting in a more demanding task than medial or coda mispronunciations.

A second question is whether the experimental context modulates infants' responses to mispronunciations. In order to study the influence of mispronunciation position, many studies control the *phonological overlap between target and distractor labels*. For example, when examining sensitivity to a vowel mispronunciation of the target word "dog", the image of a dog would be paired with a distractor image that shares onset overlap, such as "duck". This ensures that infants can not use the onset of the word to differentiate between the target and distractor images (Fernald, Swingley, & Pinto, 2001). Instead, infants must pay attention to the mispronounced phoneme in order to successfully detect the change. Most studies present infants with pictures of two name-known objects, thus ruling out the unlabeled competitor, or distractor, as possible target. In contrast, other studies present infants with pairs of familiar (labeled target) and unfamiliar (unlabeled distractor) objects (Mani & Plunkett, 2011; Skoruppa, Mani, Plunkett, Cabrol, & Peperkamp, 2013; Swingley, 2016; White & Morgan, 2008). By using an unfamiliar object as a distractor, the infant is presented with a viable option onto which the mispronounced label can be applied (Halberda, 2003; Markman, Wasow, & Hansen, 2003). This ability is developing from 18 to 30 months (Bion, Borovsky, & Fernald, 2013) and we may find that if mispronunciation sensitivity changes as children develop, that this change is modulated by *distractor familiarity*: whether the distractor used is familiar or unfamiliar. This is a particularly fruitful question to investigate within the context of a meta-analysis, as mispronunciation sensitivity in the presence of a familiar compared to unfamiliar distractor has not been directly compared.

Finally, mispronunciation sensitivity in infants has been examined in many different languages, such as English, Spanish, French, Dutch, German, Catalan, Danish, and Mandarin Chinese (see Table 1). Infants learning different languages have different ages of acquisition for words in their early lexicon, leaving direct comparisons between languages within the same study difficult and as a result rare. Indeed, the majority of studies focus on infants learning English as a native language and a sufficient sample of other languages is therefore lacking. Although we do not explicitly compare overall mispronunciation sensitivity by language (although we examine consonant and vowel mispronunciation sensitivity by language family), we assess evidence of mispronunciation sensitivity from many different languages using a meta-analytic approach.

In sum, the studies we have reviewed begin to paint a picture of the development of infants' use of phonological detail in familiar word recognition. Each study contributes one separate brushstroke and it is only by examining all of them together that we can achieve a better understanding of the big picture of early phono-lexical development. Meta-analyses can provide unique insights by estimating the population effect, both of infants' responses to correct and mispronounced labels, and of their mispronunciation sensitivity. Because we aggregate data over various age groups, this meta-analysis can also investigate the role of maturation by assessing the impact of age and vocabulary size. We also explore the influence of different linguistic (mispronunciation size, position, and type) and experimental (overlap between target and distractor labels; distractor familiarity) demands on the study of mispronunciation sensitivity. Finally, we explore potential data analysis choices that may influence different conclusions about mispronunciation sensitivity development as well as offer recommendations for experiment planning, for example by providing an effect size estimate for a priori power analyses (Bergmann et al., 2018).

## Methods

The present meta-analysis was conducted with maximal transparency and reproducibility in mind. To this end, we provide all data and analysis scripts on the supplementary website (https://osf.io/rvbjs/) and open our meta-analysis up for updates (Tsuji, Bergmann, & Cristia, 2014). The most recent version is available via the website and the interactive platform MetaLab (https://metalab.stanford.edu; Bergmann et al., 2018). Since the present paper was written with embedded analysis scripts in R (R Core Team, 2018) using the papaja package (Aust & Barth, 2018) in R Markdown (Allaire et al., 2018), it is always possible to re-analyze an updated dataset. In addition, we followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines and make the corresponding information available as supplementary materials (Moher, Liberati, Tetzlaff, Altman, & Group, 2009). Figure 1 plots our PRISMA flowchart illustrating the paper selection procedure.

**(Insert Figure 1 about here)**

## Study Selection

We first generated a list of potentially relevant items to be included in our meta-analysis by creating an expert list. This process yielded 110 items. We then used the google scholar search engine to search for papers citing the original Swingley and Aslin (2000) publication. This search was conducted on 22 September, 2017 and yielded 288 results. We removed 99 duplicate items and screened the remaining 299 items for their title and abstract to determine whether each met the following inclusion criteria: (1) original data was reported; (2) the experiment examined familiar word recognition and mispronunciations; (3) infants studied were under 31-months-of-age and typically developing; (4) the dependent variable was derived from proportion of looks to a target image versus a distractor in a eye movement experiment; (5) the stimuli were auditory speech. The final sample ($n = 32$) consisted of 27 journal articles, 1 proceedings paper, 2 theses, and 2 unpublished reports. We will refer to these items collectively as papers. Table 1 provides an overview of all papers included in the present meta-analysis.

**(Insert Table 1 about here)**

## Data Entry

The 32 papers we identified as relevant were then coded with as much consistently reported detail as possible (Bergmann et al., 2018; Tsuji et al., 2014). For each experiment (note that a paper typically has multiple experiments), we entered variables describing the publication, population, experiment design and stimuli, and results. For the planned analyses to evaluate the development of mispronunciation sensitivity and modulating factors, we focus on the following characteristics:

1. Condition: Were words mispronounced or not;
2. Mean age reported per group of infants, in days;
3. Vocabulary size, measured by a standardized questionnaire or list;
4. Position of mispronunciation: onset, medial, offset, or mixed;
5. Size of mispronunciation, measured in features changed;
6. Phonological overlap between target and distractor: onset, onset/medial, rhyme, none, novel word;
7. Type of mispronunciation: consonant, vowel, or both.
8. Distractor familiarity: familiar or unfamiliar;

A detailed explanation for moderating factors 3-8 can be found in their respective sections in the Results.[1] We separated conditions according to whether or not the target word was mispronounced to be able to investigate infants' looking to the target picture as well as their mispronunciation sensitivity, which is the difference between looks to the target in correct and mispronounced trials. When the same infants were further exposed to

[1] Two papers tested bilingual infants (Ramon-Casas & Bosch, 2010; Ramon-Casas, Swingley, Sebastián-Gallés, & Bosch, 2009), yielding 2 and 4 records. Due to this small number, we do not investigate the role of multilingualism, but do note that removing these papers from the meta-analysis did not alter the pattern of results.

multiple mispronunciation conditions and the results were reported separately in the paper, we also entered each condition as a separate row (e.g., consonant versus vowel mispronunciations; Mani & Plunkett, 2007). The fact that the same infants contributed data to multiple rows (minimally those containing information on correct and mispronounced trials) leads to shared variance across effect sizes, which we account for in our analyses (see next section). We will call each row a record; in total there were 251 records in our data.

## Data analysis

Effect sizes are reported for infants' looks to target pictures after hearing a correctly pronounced or a mispronounced label (object identification) as well as the difference between effect sizes for correct and mispronounced trials (i.e. mispronunciation sensitivity). The effect size reported in the present paper is based on comparison of means, standardized by their variance. The most well-known effect size from this group is Cohen's $d$ (Cohen, 1988). To correct for the small sample sizes common in infant research, however, we used Hedges' $g$ instead of Cohen's $d$ (Hedges, 1981; Morris & DeShon, 2002).

We calculated Hedges' $g$ using the raw means and standard deviations reported in the paper ($n$ = 177 records from 25 papers) or reported t-values ($n$ = 74 records from 9 papers). Two papers reported raw means and standard deviations for some records and just t-values for the remaining records (Altvater-Mackensen et al., 2014; Swingley, 2016). Raw means and standard deviations were extracted from figures for 3 papers. In a within-participant design, when two means are compared (i.e. looking during pre- and post-naming) it is necessary to obtain correlations between the two measurements at the participant level to calculate effect sizes and effect size variance. Upon request we were provided with correlation values for one paper (Altvater-Mackensen, 2010); we were able to compute correlations using means, standard deviations, and t-values for 5 papers (following Csibra, Hernik, Mascaro, Tatone, & Lengyel, 2016; see also Rabagliati, Ferguson, &

Lew-Williams, 2018). Correlations were imputed for the remaining papers (see Black & Bergmann, 2017 for the same procedure). For two papers, we could not derive any effect size (Ballem & Plunkett, 2005; Renner, 2017), and for a third paper, we do not have sufficient information in one record to compute effect sizes (Skoruppa et al., 2013). We compute a total of 106 effect sizes for correct pronunciations and 150 for mispronunciations. Following standard meta-analytic practice, we remove outliers, i.e. effect sizes more than 3 standard deviations from the respective mean effect size. This leads to the exclusion of 2 records for correct pronunciations and 3 records for mispronunciations.

To take into account the fact that the same infants contributed to multiple datapoints, we analyze our results in a multilevel approach using the R (R Core Team, 2018) package metafor (Viechtbauer, 2010). We use a multilevel random effects model which estimates the mean and variance of effect sizes sampled from an assumed distribution of effect sizes. In the random effect structure we take into account the shared variance of effect sizes drawn from the same paper, and nested therein that the same infants might contribute to multiple effect sizes.

Mispronunciation sensitivity studies typically examine infants' proportion of target looks (PTL) in comparison to some baseline measurement. PTL is calculated by dividing the percentage of looks to the target by the total percentage of looks to both the target and distractor images. Across papers the baseline comparison varied; since other options were not available to us, we used the baseline reported by the authors of each paper. Most papers ($n$ = 52 records from 13 papers) subtracted the PTL score for a pre-naming baseline phase from the PTL score for a post-naming phase and report a difference score.

Other papers either compared post- and pre-naming PTL with one another ($n$ = 29 records from 10 papers), thus reporting two variables, or compared post-naming PTL with a chance level of 50% ($n$ = 23 records from 9 papers). For all these comparisons, positive values (either as reported or after subtraction of chance level or a pre-naming baseline PTL) indicate target looks towards the target object after hearing the label, i.e. a recognition effect. Standardized effect sizes based on mean differences, as calculated here, preserve the sign. Consequently, positive effect sizes reflect more looks to the target picture after naming, and larger positive effect sizes indicate comparatively more looks to the target.

## Publication Bias

In the psychological sciences, there is a documented reluctance to publish null results. As a result, significant results tend to be over-reported and thus might be over-represented in our meta-analyses (see Ferguson & Heene, 2012). To examine whether this is also the case in the mispronunciation sensitivity literature, which would bias the data analyzed in this meta-analysis, we conducted two tests. We first examined whether effect sizes are distributed as expected based on sampling error using the rank correlation test of funnel plot asymmetry with the R (R Core Team, 2018) package metafor (Viechtbauer, 2010). Effect sizes with low variance were expected to fall closer to the estimated mean, while effect sizes with high variance should show an increased, evenly-distributed spread around the estimated mean. Publication bias would lead to an uneven spread.

Second, we analyze all of the significant results in the dataset using a p-curve from the p-curve app (v4.0, http://p-curve.com; Simonsohn, Nelson, & Simmons, 2014). This p-curve tests for evidential value by examining whether the p-values follow the expected distribution of a right skew in case the alternative hypothesis is true, versus a flat distribution that speaks for no effect being present in the population and all observed significant effects being spurious.

Responses to correctly pronounced and mispronounced labels were predicted to show different patterns of looking behavior. In other words, there is an expectation that infants should look to the target when hearing a correct pronunciation, but studies vary in their report of significant looks to the target when hearing a mispronounced label (i.e. there might be no effect present in the population); as a result, we conducted these two analyses to assess publication bias separately for both conditions.

### Meta-analysis

The models reported here are multilevel random-effects models of variance-weighted effect sizes, which we computed with the R (R Core Team, 2018) package metafor (Viechtbauer, 2010). To investigate how development impacts mispronunciation sensitivity, our core theoretical question, we first introduced age (centered; continuous and measured in days but transformed into months for ease of interpreting estimates by dividing by 30.44) as a moderator to our main model. Second, we analyzed the correlation between reported vocabulary size and mispronunciation sensitivity using the R (R Core Team, 2018) package meta (Schwarzer, 2007). For a subsequent investigation of experimental characteristics, we introduced each as a moderator: size of mispronunciation, position of mispronunciation, phonological overlap between target and distractor labels, type of mispronunciation, and distractor familiarity (more detail below).

## Results

### Publication Bias

Figure 2 shows the funnel plots for both correct pronunciations and mispronunciations (code adapted from Sakaluk, 2016). Funnel plot asymmetry was significant for both correct pronunciations (Kendall's $\tau = 0.53$, $p < .001$) and mispronunciations (Kendall's $\tau = 0.16$, $p = 0.004$). These results, quantifying the asymmetry in the funnel plots (Figure 2), indicate bias in the literature. This is particularly evident for correct pronunciations, where larger effect sizes have greater variance (bottom right corner) and the more precise effect sizes (i.e. smaller variance) tend to be smaller than expected (top left, outside the triangle).

The stronger publication bias for correct pronunciation might reflect the status of this condition as a control. If infants were not looking to the target picture after hearing the correct label, the overall experiment design is called into question. However, even in a well-powered study one would expect the regular occurrence of null results even though as a population infants would reliably show the expected object identification effect.

We should also point out that funnel plot asymmetry can be caused by multiple factors besides publication bias, such as heterogeneity in the data. There are various possible sources of heterogeneity, which our subsequent moderator analyses will begin to address. Nonetheless, we will remain cautious in our interpretation of our findings and hope that an open dataset which can be expanded by the community will attract previously unpublished null results so we can better understand infants' developing mispronunciation sensitivity.

**(Insert Figure 2 about here)**

We next examined the p-curves for significant values from the correctly pronounced and mispronounced conditions. The p-curve based on 72 statistically significant values for correct pronunciations indicates that the data contain evidential value (Z = -17.93, $p < .001$) and we find no evidence of a large proportion of p-values just below the typical alpha threshold of .05 that researchers consistently apply in this line of research. The p-curve based on 36 statistically significant values for mispronunciations indicates that the data contain evidential value (Z = -6.81, $p < .001$) and there is again no evidence of a large proportion of p-values just below the typical alpha threshold of .05.

Taken together, the results suggest a tendency in the literature towards publication bias. As a result, our meta-analysis may systematically overestimate effect sizes and we therefore interpret all estimates with caution. Yet, the p-curve analysis suggests that the literature contains evidential value, reflecting a "real" effect. We therefore continue our meta-analysis.

## Meta-analysis

**Object Identification for Correct and Mispronounced Words.** We first calculated the meta-analytic effect for infants' ability to identify objects when hearing correctly pronounced labels. The variance-weighted meta-analytic effect size Hedges' $g$ was 0.916 (SE = 0.122) which was significantly different from zero (CI [0.676, 1.156], $p < .001$). This is a small to medium effect size (according to the criteria set by Mills-Smith, Spangler, Panneton, & Fritz, 2015). That the effect size is significantly above zero suggests that when presented with the correctly pronounced label, infants tended to fixate on the corresponding object. Although the publication bias present in our analysis of funnel plot asymmetry suggests that the effect size Hedges' $g$ may be overestimated for object identification in response to correctly pronounced words, the p-curve results and a CI lower bound of 0.68, which is substantially above zero, together suggest that this result is somewhat robust. In other words, we are confident that the true population mean lies above zero for object recognition of correctly pronounced words.

We then calculated the meta-analytic effect for object identification in response to mispronounced words. In this case, the variance-weighted meta-analytic effect size Hedges' $g$ was 0.249 (SE = 0.06) which was also significantly different from zero (CI [0.132, 0.366], $p < .001$). This is considered a small effect size (Mills-Smith et al., 2015), but significantly above zero, which suggests that even when presented with a mispronounced label, infants fixated the correct object. In other words, infants are able to resolve mispronunciations, a key skill in language processing We again note the publication bias (which was smaller in this condition), and the possibility that the effect size Hedges' $g$ may be overestimated. But, as the p-curve indicated evidential value, we are confident in the overall pattern, namely that infants fixate the target even after hearing a mispronounced label.

**Mispronunciation Sensitivity Meta-Analytic Effect.** The above two analyses considered the data from mispronounced and correctly pronounced words separately. To evaluate mispronunciation sensitivity, we compared the effect size Hedges' $g$ for correct pronunciations with mispronunciations directly. To this end, we combined the two datasets. When condition was included (correct, mispronounced), the moderator test was significant (QM(1) = 103.408, $p < .001$). The estimate for mispronunciation sensitivity was 0.608 (SE = 0.06), and infants' looking behavior across conditions was significantly different (CI [0.49, 0.725], $p < .001$). This confirms that although infants fixate the correct object for both correct pronunciations and mispronunciations, the observed fixations to target (as measured by the effect sizes) were significantly greater for correct pronunciations. In other words, we observe a significant difference between the two conditions and can now quantify the modulation of fixation behavior in terms of standardized effect sizes and their variance. This first result has both theoretical and practical implications, as we can now reason about the amount of perturbation caused by mispronunciations and can plan future studies to further investigate this effect with suitable power.

Heterogeneity was significant for both correctly pronounced (Q(103) = 625.63, $p < .001$) and mispronounced words, (Q(146) = 462.51, $p < .001$), as well as mispronunciation sensitivity, which included the moderator condition (QE(249) = 1,088.14, $p < .001$). This indicated that the sample contains unexplained variance leading to significant difference between studies beyond what is to be expected based on random sampling error. We therefore continue with our moderator analysis to investigate possible sources of this variance.

**Object Recognition and Mispronunciation Sensitivity Modulated by Age.** To evaluate the different predictions we laid out in the introduction for how mispronunciation sensitivity will change as infants develop, we next added the moderator age (centered; continuous and measured in days but transformed into months for ease of interpreting estimates by dividing by 30.44 for Figure 3).

In the first analyses, we investigate the impact of age separately on conditions where words were either pronounced correctly or not. Age did not significantly modulate object identification in response to correctly pronounced (QM(1) = 0.558, $p = 0.455$) or mispronounced words (QM(1) = 1.64, $p = 0.2$). The lack of a significant modulation together with the small estimates for age (correct: $\beta = 0.014$, SE = 0.019, 95% CI[-0.022, 0.05], $p = 0.455$; mispronunciation: $\beta = 0.015$, SE = 0.011, 95% CI[-0.008, 0.037], $p = 0.2$) indicates that there might be no relationship between age and target looks in response to a correctly pronounced or mispronounced label. We note that the estimates in both cases are positive, however, which is in line with the general assumption that infants' language processing overall improves as they mature (Fernald, Pinto, Swingley, Weinberg, & McRoberts, 1998). We plot both object recognition and mispronunciation sensitivity as a function of age in Figure 3.

We then examined the interaction between age and mispronunciation sensitivity (correct vs. mispronounced words) in our whole dataset. The moderator test was significant (QM(3) = 106.158, $p < .001$). The interaction between age and mispronunciation sensitivity, however, was not significant ($\beta = 0.012$, SE = 0.013, 95% CI[-0.014, 0.039], $p = 0.349$); the moderator test was mainly driven by the difference between conditions. The small estimate, as well as inspection of Figure 3, suggests that as infants age, their mispronunciation sensitivity neither increases or decreases.

**(Insert Figure 3 about here)**

**Vocabulary Size: Correlation Between Mispronunciation Sensitivity and Vocabulary.** Of the 32 papers included in the meta-analysis, 13 analyzed the relationship between vocabulary scores and object recognition for correct pronunciations and mispronunciations (comprehension = 11 papers and 39 records; production = 3 papers and 20 records). Children comprehend more words than they can produce, leading to different estimates for comprehension and production. Production data is easier to estimate for parents in the typical questionnaire-based assessment and may therefore be more reliable (Tomasello & Mervis, 1994). As a result, we planned to analyze these two types of vocabulary measurement separately. However, because only 3 papers reported correlations with productive vocabulary scores, only limited conclusions can be drawn. We also note that because individual effect sizes in our analysis were related to object recognition and not mispronunciation sensitivity, we were only able to calculate the relationship between vocabulary scores and the former. In our vocabulary analysis, we therefore focus exclusively on the relationship between comprehension and object recognition for correct pronunciations and mispronunciations.

We first considered the relationship between vocabulary and object recognition for correct pronunciations. Higher comprehension scores were associated with greater object recognition in response to correct pronunciations for 9 of 10 records, with correlation values ranging from -0.16 to 0.48. The weighted mean effect size Pearson's $r$ of 0.14 was small but did differ significantly from zero (CI [0.03; 0.25] $p = 0.012$). As a result, we can draw a tentative conclusion that there is a positive relationship between comprehension scores and object recognition in response to correct pronunciations.

We next considered the relationship between vocabulary and object recognition for mispronunciations. Higher comprehension scores were associated with greater object recognition in response to mispronunciations for 17 of 29 records, with correlation values ranging from -0.35 to 0.57. The weighted mean effect size Pearson's $r$ of 0.05 was small and did not differ significantly from zero (CI [-0.01; 0.12] $p = 0.119$). The small correlation suggests either a very small positive or no relationship between vocabulary and object recognition for mispronunciations.

Figure 4 plots the year of publication for all the mispronunciation sensitivity studies included in this meta-analysis. This figure illustrates two things: the increasing number of mispronunciation sensitivity studies in general and the decreasing number of mispronunciation studies measuring vocabulary. The lack of evidence for a relationship between mispronunciation sensitivity and vocabulary size in some early studies may have contributed to increasingly fewer researchers including vocabulary measurements in their mispronunciation sensitivity experimental design. This may explain our underpowered analysis of the relationship between object recognition for correct pronunciations and mispronunciations and vocabulary size, despite its theoretical interest.

**(Insert Figure @ref(fig:Vocabdescribe1 about here)**

**Interim discussion: Development of infants' mispronunciation sensitivity.** The main goal of this paper was to assess mispronunciation sensitivity and its maturation with age and increased vocabulary size. The results seem clear: Although infants consider a mispronunciation to be a better match to the target image than to a distractor image, there was a constant and stable effect of mispronunciation sensitivity across all ages. In the literature, evidence for all possible developmental trajectories has been found, including mispronunciation sensitivity that increases, decreases, or does not change with age or vocabulary size. The present results do lend some support for the proposal that mispronunciation sensitivity stays constant as infants develop. Furthermore, although we found a relationship between vocabulary size (comprehension) and target looking for correct pronunciations, we found no relationship between vocabulary and target looking for mispronunciations. This may be due to too few studies including reports of vocabulary size and more investigation is needed to draw a firm conclusion.

The studies examined in this meta-analysis examined mispronunciation sensitivity, but many also included more specific questions aimed at uncovering more detailed phonological processes at play during word recognition. Not only are these questions theoretically interesting, they also have the potential to change the difficulty of a msipronunciation sensitivity experiment. It is possible that the lack of developmental change in mispronunciation sensitivity found by our meta-analysis does not capture a true lack of change, but is instead influenced by differences in the types of tasks given to infants of different ages. If infants' word recognition skills are generally thought to improve with age and vocabulary size, research questions that tap more complex processes may be more likely to be investigated in older infants. In the following section, we investigate the role that different moderators play in mispronunciation sensitivity. To investigate the possibility of systematic differences in the tasks across ages, we additionally include an exploratory analysis of whether different moderators and experimental design features were included at different ages.

## Moderator Analyses

In this section, we consider each moderator individually and investigate its influence on mispronunciation sensitivity. For most moderators (except mispronunciation size), we combine the correct and mispronounced datasets and include the moderator of condition, to study mispronunciation sensitivity as opposed to object recognition. To better understand the impact of these moderators on developmental change, we include age as subsequent moderator. Finally, we analyze the relationship between infant age and the moderator condition they were tested in using Fisher's exact test, which is more appropriate for small sample sizes (Fisher, 1922). This evaluates the independence of infants' age group (divided into quartiles unless otherwise specified) and assignment to each type of condition in a particular moderator.

**Size of mispronunciation.** To assess whether the size of the mispornunciation tested, as measured by the number of features changed, modulates mispronunciation sensitivity, we calculated the meta-analytic effect for object identification in response to words that were pronounced correctly and mispronounced using 1-, 2-, and 3-feature changes. We did not include data for which the number of features changed in a mispronunciation was not specified or the number of features changed was not consistent (e.g., one mispronunciation included a 2-feature change whereas another only a 1-feature change). This analysis was therefore based on a subset of the overall dataset, with 90 records for correct pronunciations, 99 for 1-feature mispronunciations, 16 for 2-feature mispronunciations, and 6 for 3-feature mispronunciations. Each feature change (from 0 to 3; 0 representing correct pronunciations) was considered to have an equal impact on mispronunciation sensitivity, following the argument of graded sensitivity (Mani & Plunkett, 2011; White & Morgan, 2008), and this modertor was coded as a continuous variable.

To understand the relationship between mispronunciation size and mispronunciation sensitivity, we evaluated the effect size Hedges' $g$ with number of features changed as a moderator. The moderator test was significant, QM(1) = 61.081, $p < .001$. Hedges' $g$ for number of features changed was -0.406 (SE = 0.052), which indicated that as the number of features changed increased, the effect size Hedges' $g$ significantly decreased (CI [-0.507, -0.304], $p < .001$). We plot this relationship in Figure 5. This confirms previous findings of a graded sensitivity to the number of features changed for both consonant (Bernier & White, 2017; Tamasi, 2016; White & Morgan, 2008) and vowel (Mani & Plunkett, 2011) mispronunciations as well as the importance of controlling for the degree of phonological mismatch in experimental design. In other words, the infants' ability to detect a mispronunciation depends on the size of the mispronunciation.

When age was added as a moderator to the model, the moderator test was significant, QM(3) = 143.617, $p < .001$, but the estimate for the interaction between age and number of features changed was small and not significant, $\beta = 0.009$, SE = 0.006, 95% CI[-0.002, 0.02], $p = 0.099$. This suggests that the impact of number of features changed on mispronunciation sensitivity does not substantially change with infant age. We note, however, that only a handful of studies have explicitly examined the effect of the number of features changed on mispronunciation sensitivity and only these studies include 3-feature changes (Bernier & White, 2017; Mani & Plunkett, 2011; Tamasi, 2016; White & Morgan, 2008), which may narrow our ability to draw conclusions about developmental change.

Finally, results of Fisher's exact test were not significant, $p = 0.703$. This lack of a relationship suggests that older and younger infants are not being tested in experimental conditions that differentially manipulate the number of features changed.

**(Insert Figure 5 about here)**

**Position of mispronunciation.** We next calculated the meta-analytic effect of mispronunciation sensitivity (moderator: condition) in response to mispronunciations on the onset, medial, and coda phonemes. We did not include data for which the mispronunciation varied within record in regard to position ($n = 40$) or was not reported ($n = 10$). The analysis was therefore based on a subset of records of the overall dataset, testing mispronunciations on the onset ($n = 143$ records), medial ($n = 48$), and coda ($n = 10$) phonemes. We coded the onset, medial, and coda positions as continuous variables, to evaluate the importance of each subsequent position (Marslen-Wilson & Zwitserlood, 1989).

When mispronunciation position was included as a moderator, the moderator test was significant, QM(3) = 172.345, $p < .001$. For the interaction between condition and mispronunciation position, the estimate was small but significant ($\beta$ = -0.126, SE = 0.064, 95% CI[-0.252, 0], $p = 0.049$. As can be seen in Figure 6, mispronunciation sensitivity decreased linearly as the position of the mispronunciation moved later in the word, with sensitivity greatest for onset mispronunciations and smallest for coda mispronunciations.

When age was added as a moderator, the moderator test was significant, QM(7) = 175.856, $p < .001$. The estimate for the three-way interaction between age, condition, and mispronunciation position was small and not significant ($\beta$ = 0.022, SE = 0.018, 95% CI[-0.013, 0.057], $p = 0.223$.

Due to the small sample size of coda mispronunciations, we only included 3 age groups in Fisher's exact test. The results were significant, $p = 0.02$. Older infants were more likely to be tested on onset mispronunciations, while younger infants were more likely to be tested on medial mispronunciations. An onset mispronunciation may be more disruptive to lexical access than mispronunciations in subsequent positions (Marslen-Wilson & Zwitserlood, 1989), and therefore easier to detect. For this reason, it is rather unsuprising that onset mispronunciations show the greatest estimate of mispronunciation sensitivity. However, it also means that younger infants, who were more likely to be tested on medial mispronunciations, had a comparably harder task than older infants, who were more likely to be tested on onset mispronunciations. It is unlikely that this influenced our developmental trajectory estimate, as the consequence would have been mispronunciation sensitivity that increases with age.

**(Insert Figure 6 about here)**

**Type of mispronunciation (consonant or vowel).** We next calculated the meta-analytic effect of mispronunciation sensitivity (moderator: condition) in response to the type of mispronunciation, consonant or vowel. Furthermore, sensitivity to consonant and vowel mispronunciations is hypothesized to differ depending on whether the infant is learning a Germanic or Romance language. We therefore conducted two sets of analyses, one analyzing consonants and vowels alone and a second including langauge family as a moderator. We did not include data for which mispronunciation type varied within experiment and was not reported separately ($n$ = 23). The analysis was therefore based on a subset of the overall dataset, comparing records with consonant ($n$ = 145) and vowel ($n$ = 71) mispronunciations.

When mispronunciation type was included as a moderator, the moderator test was significant, QM(7) = 153.795, $p < .001$, but the interaction between mispronunciation type and condition ($\beta$ = 0.056, SE = 0.079, 95% CI[-0.099, 0.211], $p$ = 0.479) was not significant. The results suggest that overall, infants' sensitivity to consonant and vowel mispronunciations was similar (Figure 7a).

When age was added as a moderator, the moderator test was significant, QM(7) = 153.795, $p < .001$ and the estimate for the three-way interaction between age, condition, and mispronunciation type was significant, but relatively small ($\beta = 0.044$, SE = 0.018, 95% CI[0.008, 0.08], $p = 0.016$. As can be seen in Figure 7b, as infants age, mispronunciation sensitivity grows larger for vowel mispronunciations but stays steady for consonant mispronunciations. Noticeably, mispronunciation sensitivity appears greater for consonant compared to vowel mispronunciations at younger ages, but this difference diminishes as infants age.

The results of Fisher's exact test were significant, $p < .001$. Older infants were more likely to be tested on consonant mispronunciations, while younger infants were more likely to be tested on vowel mispronunciations. It is not immediately clear whether the relationship between infant age and type of mispronunciation influences our estimate of how mispronunciation sensitivity changes with development. Whether consonant or vowel mispronunciations are more "difficult" is a matter of theoretical debate, but some evidence suggest that it may be influenced by infants' native language (Nazzi et al., 2016). We next examined whether this was the case.

**(Insert Figure 7 about here)**

We first classified infants into language families. Infants learning American English ($n = 56$), British English ($n = 66$), Danish ($n = 6$), Dutch ($n = 58$), and German ($n = 21$) were classified into the Germanic language family ($n = 207$). Infants learning Catalan ($n = 4$), Spanish ($n = 4$), French ($n = 8$), Catalan and Spanish simultaneously (i.e. bilinguals; $n = 6$), and Swiss French ($n = 6$) were classified into the Romance language family ($n = 28$).

When language family was included as a moderator, the moderator test was significant, QM(7) = 158.889, $p < .001$. The three-way interaction between mispronunciation type, condition, language family was large and also significant , $\beta =$

-0.872, SE = 0.28, 95% CI[-1.421, -0.323], $p = 0.002$. As can be seen in Figure 8a, mispronunciation sensitivity for consonants was similar for Germanic and Romance languages. Mispronunciation sensitivity for vowels, however, was greater for Germanic compared to Romance languages.

We next added age as a moderator, resulting in a significant moderator test, QM(15) = 185.148, $p < .001$, and a small but significant estimate for the four-way interaction between mispronunciation type, condition, language family, and age $\beta = 0.331$, SE = 0.078, 95% CI[0.178, 0.484], $p < .001$. As can also be seen in Figure 8b, for infants learning Germanic languages, sensitivity to consonant and vowel mispronunciations did not change with age. In contrast, infants learning Romance languages show a decrease in sensitivity to consonant mispronunciations, but an increase in sensitivity to vowel mispronunciations with age.

We were unable to use Fisher's exact test to evaluate whether infants of different ages were more or less likely to be tested on consonant or vowel mispronunciations depending on their native language. This was due to the small sample size of infants learning Romance languages ($n = 28$).

**(Insert Figure 8 about here)**

**Phonological overlap between target and distractor.** We next examined the meta-analytic effect of mispronunciation sensitivity (moderator: condition) in response to mispronunciations when the target-distractor pairs either had no overlap or shared the same onset phoneme. We did not include data for which the overlap included both the onset and medial phonemes ($n = 4$), coda phonemes ($n = 3$), or for targets paired with an unfamiliar distractor image ($n = 60$). The analysis was therefore based on a subset of the overall dataset, comparing 104 records containing onset phoneme overlap between the target and distractor with 80 containing no overlap between target and distractor.

When target-distractor overlap was included as a moderator, the moderator test was significant, QM(3) = 48.101, $p < .001$. The estimate for the interaction between condition and distractor overlap was small, but significant ($\beta = 0.195$, SE = 0.213, 95% CI[-0.223, 0.612], $p = 0.36$, suggesting that mispronunciation sensitivity was greater when target-distractor pairs shared the same onset phoneme compared to when they shared no phonological overlap. This relationship be seen in Figure 9a.

When age was added as a moderator, the moderator test was significant, QM(7) = 67.82, $p < .001$ and the estimate for the three-way interaction between age, condition, and distractor overlap was significant, but relatively small ($\beta = = 0.091$, SE = 0.038, 95% CI[0.017, 0.166], $p = 0.016$. As can be seen in Figure 9b, mispronunciation sensitivity increases with age for target-distractor pairs containing onset overlap, but decreases with age for target-distractor pairs containing no overlap.

The results of Fisher's exact test were significant, $p < .001$. Older infants were more likely to be tested in experimental conditions where target and distractor images overlapped on their onset phoneme, while younger infants were more likely to be tested with target and distractor images that did not control for overlap. A distractor image that overlaps in the onset phoneme with the target image is considered a more challenging task to the infant, as infants must pay attention to the mispronounced phoneme and can not use the differing onsets between target and distractor images to differentiate (Fernald et al., 2001). It therefore appears that older infants were given a more challenging task than younger infants. We return to this issue in the General Discussion.

**(Insert Figure 9 about here)**

**Distractor familiarity.** We next calculated the meta-analytic effect of mispronunciaiton sensitivity (moderator: condition) in experiments were the target image was paired with a familiar or unfamiliar distractor image. A familiar distractor was used in

179 records and a unfamiliar distractor in 72 records.

When distractor familiarity was included as a moderator, the moderator test was significant, QM(1) = 61.081, $p < .001$, but the effect of distractor familiarity ($\beta$ = -0.12, SE = 0.144, 95% CI[-0.403, 0.162], $p = 0.403$) as well as the interaction between distractor familiarity and condition ($\beta$ = 0.067, SE = 0.137, 95% CI[-0.203, 0.336], $p = 0.628$) were not significant. The results suggest that overall, infants' familiarity with the distractor object (familiar or unfamiliar) did not impact their mispronunciation sensitivity.

When age was added as a moderator, the moderator test was significant QM(7) = 107.683, $p < .001$. The estimate for the three-way-interaction between condition, distractor familiarity, and age was small and not significant ($\beta$ = = -0.021, SE = 0.035, 95% CI[-0.09, 0.048], $p = 0.547$. These results suggest that regardless of age, mispronunciation sensitivity was similar whether the distractor image was familiar or unfamiliar.

The results of Fisher's exact test were not significant, $p = 0.072$. This lack of a relationship suggests that older and younger infants were not tested in experimental conditions that differentially employ distractor images that are familiar or unfamiliar.

**Interim discussion: Moderator analyses.** Next to the main goal of this paper, which was to evaluate the development of infants' sensitivity to mispronunciations, we also investigated the more nuanced questions often posed in studies investigating infants' mispronunciation sensitivity. We identified five additional manipulations often present in mispronunciation sensitivity studies and investigated the how those manipulations modulated mispronunciation sensitivity and whether this changed with infant age. Furthermore, considering the lack of developmental change found in our main analysis, we evaluted whether these additional manipulations were disproportionatly conducted with children of different ages, to assess whether older infants recieve more difficult tasks than younger infants.

To briefly summarize, mispronunciation sensitivity was modulated overall by the size of the mispronunciation tested, whether target-distractor pairs shared phonological overlap, and the position of the mispronunciation. Neither distractor familiarity (familiar, unfamiliar) or type of mispronunciation (consonant, vowel) were found to impact mispronunciation sensitivity. The developmental trajectory of mispronunciation sensitivity was influenced by type of mispronunciation and overlap between the target and distractor labels, but mispronunciation size, mispronunciation position, and distractor familiarity were found to have no influence. Finally, in some cases there was evidence that older and younger infants were given experimental manipulations that may have rendered the experimental task more or less difficult. In one instance, younger infants were given a more difficult task, mispronunciations on the medial position, which is unlikely to contribute to the lack of developmental effects in our main analysis. Yet, his was not always the case; in a different instance, older children were more likely to be given target-distractor pairs that overlapped on their onset phoneme, a situation in which it is more difficult to detect a mispronunciation and may have bearing on our main developmental results. We return to these findings in the General Discussion.

## Exploratory Analyses

We next considered whether an effect of maturation might have been masked by other factors we have not yet captured in our analyses. A strong candidate that emerged during the construction of the present dataset and careful reading of the original papers was the analysis approach. We observed, as mentioned in the Methods section, large variation in the dependent variable reported, and additionally noted that the size of the chosen post-naming analysis window varied substantially across papers. Researchers' analysis strategy may be adapted to infants' age or influenced by having observed the data. For example, consider the possibility that there is a true increase in mispronunciation sensitivity over development. In this scenario, younger infants should show no or only little sensitivity to mispronunciations while older infants would show a large sensitivity to mispronunciations. This lack of or small mispronunciation sensitivity in younger infants is likely to lead to non-significant results, which would be more difficult to publish (Ferguson & Heene, 2012). In order to have publishable results, adjustments to the analysis approach could be made until a significant effect of mispronunciation sensitivity is found. This would lead to an increase in significant results and alter the observed developmental trajectory of mispronunciation sensitivity in the current meta-analysis. Such a scenario is in line with the publication bias we observe (Simmons, Nelson, & Simonsohn, 2011). We examine whether variation in the approach to data analysis may be have an influence on our conclusions regarding infants' developing mispronunciation sensitivity.

We analyzed analysis choices related to timing (post-naming analysis window; offset time) and type of dependent variable in our coding of the dataset because they are consistently reported and might be useful for experiment design in the future by highlighting typical choices and helping establish field standards. In the following, we discuss the possible theoretical motivation for these data analysis choices, the variation present in the current meta-analysis dataset, and the influence these analysis choices may have on measurements of mispronunciation sensitivity development. We focus specifically on the size of the mispronunciation sensitivity effect, considering the whole dataset and including condition (correct pronunciation, mispronunciation) as moderator.

**Timing.** When designing mispronunciation sensitivity studies, experimenters can choose the length of time each trial is presented. This includes both the length of time before the target object is named (pre-naming phase) as well as after (post-naming phase) and is determined prior to data collection. If the post-naming phase represents the amount of time the infant viewed the target-distractor image pairs after auditory presentation of the target word, then the post-naming analysis window represents how much of this phase was included in the statistical analysis. Unlike the post-naming phase, however, the post-naming analysis window can be chosen after the experimental data is collected. Evidence suggests that the speed of word recognition processing is slower in young infants

(Fernald et al., 1998), which may lead researchers to include longer post-naming phases in their experiments with younger infants. If this is the case, we expect a negative correlation between post-naming phase length and infant age.

Across papers, the length of the post-naming phase varied from 2000 to 9000 ms, with a median value of 3500 ms. The most popular post-naming phase length was 4000 ms, used in 74 records. Regarding the post-naming analysis window, half of the records were analyzed using the whole post-naming phase presented to the infant ($n$ = 124), while the other half were analyzed using a shorter portion of the post-naming time window, usually excluding later portions ($n$ = 127). Across papers, the length of the post-naming analysis window varied from 1510 to 4000 ms, with a median value of 2500 ms. The most popular post-naming analysis window length was 2000 ms, used in 97 records.

There was no apparent relation between infant age and post-naming phase length ($r$ = 0.01, 95% CI[-0.11, 0.13], $p$ = 0.882), but there was a significant negative relationship between infant age and post-naming analysis window length, such that younger infants' looking times were analyzed using a longer post-naming analysis window ($r$ = -0.23, 95% CI[-0.35, -0.11], $p < .001$). Although we observe no relationship between age and post-naming phase length, a value that is determine before data collection, we do observe a relationship with post-naming analysis window length, a value that may be determined after data collection and even driven by observation of the data itself. In other words, we observe variation in time-related analysis decisions related to infants' age.

Another potential source of variation considers the amount of time it takes for an eye movement to be initiated in response to a visual stimulus, which we refer to as offset time (time between the onset of the target word and the offset of the post-naming analysis window). Previous studies examining simple stimulus response latencies first determined that infants require at least 233 ms to initiate an eye-movement in response to a stimulus (Canfield & Haith, 1991). In the first infant mispronunciation sensitivity study, Swingley and Aslin (2000) used an offset time of 367 ms, which was "an 'educated guess' based on studies... showing that target and distractor fixations tend to diverge at around 400 ms." (Swingley & Aslin, 2000, p. 155). Upon inspecting the offset time values used in the papers in our meta-analysis, the majority used a similar offset time value (between 360 and 370 ms) for analysis ($n$ = 151), but offset values ranged from 0 to 500 ms, and were not reported for 36 records. We note that Swingley (2009) also included offset values of 1133 ms to analyze responses to coda mispronunciations. There was an inverse relationship between infant age and size of offset, such that younger infants were given longer offsets, although this correlation was not significant ($r$ = -0.10, 95% CI[-0.23, 0.03], $p$ = 0.13). This lack of a relationship is possibly driven by the field's consensus that an offset of about 367 ms is appropriate for analyzing word recognition in infants, including studies that evaluate mispronunciation sensitivity.

Although there are a priori reasons, such as infant age or previous studies, to choose the post-naming analysis window or offset time, these choices may occur after data collection and might therefore lead to a higher rate of false-positives (Gelman & Loken, 2013). Considering that these choices were systematically different across infant ages, at least for the post-naming analysis window, we next explored whether the post-naming analysis window length or the offset time influenced our estimate of infants' sensitivity to mispronunciations.

***Post-naming analysis window length.***

We first assessed whether post-naming analysis window length had an impact on the overall size of the reported mispronunciation sensitivity. We considered data from both conditions in a joint analysis and included condition (correct pronunciation, mispronunciation) as an additional moderator. The moderator test was significant (QM(3) = 236.958, $p < .001$). The estimate for the interaction between post-naming analysis window and condition was small but significant ($\beta$ = -0.262, SE = 0.059, 95% CI[-0.377, -0.148], $p < .001$). This relationship is plotted in Figure 10a. These results show that as the length of the post-naming analysis window increased, the difference between target fixations for correctly pronounced and mispronounced items (mispronunciation sensitivity) decreased.

Considering that we found a significant relationship between the post-naming analysis window length and infant age, such that younger ages had a longer window of analysis, we next examined whether post-naming analysis window length modulated the estimated size of mispronunciation sensitivity as infant age changed. When age was included as a moderator, the moderator test was significant (QM(7) = 247.322, $p < .001$). The estimate for the three-way-interaction between condition, post-naming analysis window, and age was small, but significant ($\beta$ = -0.04, SE = 0.014, 95% CI[-0.068, -0.012], $p = 0.006$). As can be seen in Figure 10b, when records were analyzed with a post-naming analysis window of 2000 ms or less, mispronunciation sensitivity seems to increase with infant age. If the post-naming analysis window is greater than 2000 ms, however, there is no or a negative relation between mispronunciation sensitivity and age. In other words, all three possible developmental hypotheses might be supported depending on analysis choices made regarding post-naming analysis window length. These results suggest that conclusions about the relationship between infant age and mispronunciation sensitivity may be mediated by the size of the post-naming analysis window.

**(Insert Figure 10 about here)**

#### ***Offset time after target naming.***

We next assessed whether offset time had an impact on the size of the reported mispronunciation sensitivity. When we included both condition and offset time as moderators, the moderator test was significant (QM(3) = 236.958, $p < .001$), but the estimate for the interaction between offset time and condition was zero ($\beta = 0$, SE = 0, 95% CI[-0.001, 0], $p = 0.505$). Although we found no relationship between offset time and infant age, we also examined whether the size of offset time modulated the measure of mispronunciation sensitivity over infant age. When both offset time and condition were included as moderators, the moderator test was significant (QM(7) = 200.867, $p < .001$), but the three-way-interaction between condition, offset time, and age was again zero ($\beta =$ 0, SE = 0, 95% CI[0, 0], $p = 0.605$). Taken together, these results suggest that offset time does not modulate measured mispronunciation sensitivity nor its developmental trajectory.

## Dependent variable

Mispronunciation sensitivity experiments typically include a phase where a naming event has not yet occurred (pre-naming phase). This is followed by a naming event, whether correctly pronounced or mispronounced, and the subsequent phase (post-naming phase). The purpose of the pre-naming phase is to ensure that infants do not have systematic preferences for the target or distractor (greater interest in a cat compared to a cup) which may add variance to PTL scores in the post-naming phase. As described in the Methods section, however, there was considerable variation across papers in whether this pre-naming phase was used as a baseline measurement, or whether a different baseline measurement was used. This resulted in different measured outcomes or dependent variables. Over half of the records ($n$ = 129) subtracted the PTL score for a pre-naming phase from the PTL score for a post-naming phase, resulting in a Difference Score. The Difference Score is one value, which is then compared with a chance value of 0. In contrast, Pre vs. Post ($n$ = 69 records), directly compare the post- and pre-naming PTL scores with one another using a statistical test (e.g. t-test, ANOVA). This requires two values, one for the pre-naming phase and one for the post-naming phase. A positive Difference Score or a greater post compared to pre-naming phase PTL indicates that infants increased their target looks after hearing the naming label. The remaining records used a Post dependent variable ($n$ = 53 records), which compares the post-naming PTL score with a chance value of 50%. Here, the infants' pre-naming phase baseline preferences are not considered and instead target fixations are evaluated based on the likelihood to fixate one of two pictures (50%). As most papers do not specify whether these calculations are made before or after aggregating across trials, we make no assumptions about when this step is taken.

The Difference Score and Pre vs. Post can be considered similar to one another, in that they are calculated on the same type of data and consider pre-naming preferences. The Post dependent variable, in contrast, does not consider pre-naming baseline preferences. To our knowledge, there is no theory or evidence that explicitly drives choice of dependent variable in analysis of preferential looking studies, which may explain the wide variation in dependent variable reported in the papers included in this meta-analysis. We next explored whether the type of dependent variable calculated influenced the estimated size of sensitivity to mispronunciations. Considering that the dependent variable Post differs in its consideration of pre-naming baseline preferences, substituting these for a chance value, we directly compared mispronunciation sensitivity between Post as a reference condition and both Difference Score and Pre vs. Post dependent variables.

When we included both condition and dependent variable as moderators, the moderator test was significant (QM(5) = 259.817, $p < .001$). The estimate for the interaction between Pre vs. Post and condition was significantly smaller than that of the Post dependent variable ($\beta$ = -0.392, SE = 0.101, 95% CI[-0.59, -0.194], $p < .001$), but the difference between the Difference Score and Post in the interaction with condition was small and not significant ($\beta$ = -0.01, SE = 0.098, 95% CI[-0.203, 0.183], $p$ = 0.916). This relationship is plotted in Figure 11a. The results suggest that the reported dependent variable significantly impacted the size of the estimated mispronunciation sensitivity effect, such that studies reporting the Post. vs. Pre dependent variable showed a smaller mispronunciation sensitivity effect than those reporting Post, but that there was no difference between the Difference Score and Post dependent variables.

When age was included as an additional moderator, the moderator test was significant (QM(11) = 273.585, $p < .001$). The estimate for the interaction between Pre vs. Post, condition, and age was significantly smaller than that of the Post dependent variable ($\beta$ = -0.089, SE = 0.03, 95% CI[-0.148, -0.03], $p$ = 0.003), but the difference between the Difference Score and Post in the interaction with condition and age was small and not significant ($\beta$ = -0.036, SE = 0.027, 95% CI[-0.088, 0.016], $p$ = 0.174). When the dependent variable reported was Pre vs. Post, mispronunciation sensitivity was found to decrease with infant age, while in comparison, when the dependent variable was Post, mispronunciation sensitivity was found to increase with infant age (see This relationship is plotted in Figure 11b.)

Similar to post-naming analysis window length, all three possible developmental hypotheses might be supported depending on the dependent variable reported. In other words, choice of dependent variable may influence the conclusion drawn regarding how mispronunciation sensitivity may change with infant age. We address this issue in the General Discussion.

**(Insert Figure 11 about here)**

## General Discussion

In this meta-analysis, we set out to quantify and assess the developmental trajectory of infants' sensitivity to mispronunciations. Overall, the results of the meta-analysis showed that infants reliably fixate the target object when hearing both correctly pronounced and mispronounced labels. Infants not only recognize object labels when they were correctly pronounced, but are also likely to accept mispronunciations as labels for targets, in the presence of a distractor image. Nonetheless, there was a considerable difference in target fixations in response to correctly pronounced and mispronounced labels, suggesting that infants show an overall mispronunciation sensitivity based on the current experimental literature. In other words, infants show sensitivity to what constitutes unacceptable, possibly meaning-altering variation in word forms, thereby displaying knowledge of the role of phonemic changes throughout the ages assessed here (6 to 30 months). At the same time, infants, like adults, can recover from mispronunciations, a key skill in language processing.

Considering the variation in findings of developmental change in mispronunciation sensitivity (see Introduction), we next evaluated the developmental trajectory of infants' mispronunciation sensitivity, envisioning three possible developmental patterns: increasing, decreasing, and unchanging sensitivity. Our analysis of this relationship using age as a moderator revealed a pattern of unchanging sensitivity, which has been reported by a handful of studies directly comparing infants over a small range of ages, such as 18-24 months (Bailey & Plunkett, 2002; Swingley & Aslin, 2000) or 12-17 months (Zesiger et al., 2012). The estimated effect size for mispronunciation sensitivity in our meta-analysis suggests that sensitivity is similar across the range of 6- to 30-month-old infants tested in the studies we include. Furthermore, an examination of the influence of vocabulary size revealed no relationship between object recognition in response to mispronunciations.

Typically, vocabulary growth is thought to invoke changes in mispronunciation sensitivity. The need for phonologically well-specified word representations increases as children learn more words and must differentiate between them (Charles-Luce & Luce, 1995). Despite the theoretical implications, fewer than half of the papers included in this meta-analysis measured vocabulary ($n$ = 13; out of 32 papers total; see also Figure 4). There are more mispronunciation sensitivity studies published every year, perhaps due to the increased use of eye-trackers, which reduce the need for offline coding and thus make data collection much more efficient, but this has not translated to an increasing number of mispronunciation sensitivity studies also reporting vocabulary scores. We suggest that this may be the result of publication bias favoring significant effects or an overall hesitation to invest in data collection that is not expected to yield significant outcomes. More experimental work investigating and reporting the relationship between mispronunciation sensitivity and vocabulary size is needed if this is to be evaluated.

What do our results mean for theories of language development? Evidence that infants accept a mispronunciation (object identification) while simultaneously holding correctly pronounced and mispronounced labels as separate (mispronunciation sensitivity) may indicate an abstract understanding of words' phonological structure being in place early on. It appears that young infants may understand that the phonological form of mispronunciations and correct pronunciations do not match, but that the mispronunciation is a better label for the target compared to the distractor image. The lack of age or vocabulary effects in our meta-analysis suggest that this understanding is present from an early age and is maintained throughout early lexical development.

## Moderator Analyses

With perhaps a few exceptions, the main focus of many of the experiments included in this meta-analysis was not to evaluate whether infants are sensitive to mispronunciations in general but rather to investigate questions related to phonological and lexical processing and development. We included a set of moderator analyses to better understand these issues by themselves, as well as how they may have impacted our main investigation of infants' development of mispronunciation sensitivity. Several of these moderators include manipulations that make mispronunciation detection more or less difficult for the infant. As a result, the size of the mispronunciation sensitivity effect may be influenced by the task demands placed on the infant, especially if older infants are given more demanding tasks in comparison to younger infants, potentially masking developmental effects. Considering this, we also evaluated whether the investigation of each of these manipulations was distributed evenly across infant ages, where an uneven distribution may have subsequently heightened or dampened our estimate of developmental change.

The results of the moderator analysis reflect several findings that have been found in the literature. Although words differ from one another on many acoustic dimensions, changes in phonemes, as measured by phonological features, signal changes in meaning.

Several studies have found that infants show graded sensitivity to mispronunciations that differ in 1-, 2-, and 3-features from the correct pronunciation (Bernier & White, 2017; Mani & Plunkett, 2011; Tamasi, 2016; White & Morgan, 2008). This was also captured in our meta-analysis, which showed that for each increase in number of phonological features changed, the effect size estimate for looks to the target decreases by -0.41. Yet, this graded sensitivity appears to be stable across infant ages, although our analysis was likely underpowered. At least one study suggests that this graded sensitivity develops with age, but this was the only study to examine more than one age (Mani & Plunkett, 2011). All other studies only test one age (Bernier & White, 2017; Tamasi, 2016; White & Morgan, 2008). With more studies investigating graded sensitivity at multiple ages in infancy, we would achieve a better estimate of whether this is a stable or developing ability.

Although some theories place greater importance on onset position for word recognition and decreasing importance for phonemes in subsequent positions (i.e. COHORT; Marslen-Wilson & Zwitserlood, 1989), other theories suggest that lexical access can still recover from onset and medial mispronunciations (i.e. TRACE; McClelland & Elman, 1986). Although many studies have examined mispronunciations on multiple positions, the handful of studies that have directly directly compared sensitivity between different positions find that position of the mispronunciation does not modulate sensitivity (Swingley, 2009; Zesiger et al., 2012). This stands in contrast to the findings of our meta-analysis, which showed that for each subsequent position in the word that is changed, from onset to medial and medial to coda, the effect size estimate for looks to the target decreases by -0.13; infants are more sensitive to changes in the sounds of familiar words when they occur in an earlier position as opposed to a late position.

One potential explanation for the discrepancy between the results of individual studies and that of the current meta-analysis is the difference in how analysis timing is considered depending on the position of the mispronunciation. For example, Swingley (2009) adjusted the offset time from 367 ms for onset mispronunciations to 1133 for coda mispronunciations, to ensure that infants have a similar amount of time to respond to the mispronunciation, regardless of position. In contrast, if an experiment compares different kinds of medial mispronunciations, as in Mani and Plunkett (2011), it is not necessary to adjust offset time because the mispronunciations have a similar onset time. The length of the post-naming analysis window does impact mispronunciation sensitivity, as we discuss below, and by comparing effect sizes for different mispronuciation positions where position timing was not considered, mispronunciations that occur later in the word (i.e. medial and coda mispronunciations) may be at a disadvantage relative to onset mispronunciations. These issues can be addressed with the addition of more experiments that directly compare sensitivity to mispronunciations of different positions, as well as the use of analyses that account for timing differences.

For several moderators, we found no evidence of modulation of mispronunciation sensitivity. For example, sensitivity to mispronunciations was similar for experimental conditions that included either a familiar or an unfamiliar distrator image. Studies that include an unfamiliar, as opposed to familiar distractor image, often argue that the unfamiliar image provides a better referent candidate for mispronunciation than a familiar distractor image, where the name is already known. No studies have directly compared mispronunciation sensitivity for familiar and unfamiliar distractors, but these results suggest that this manipulation alone makes little difference in the design of the experiment. It remains possible that distractor familiarity interacts with other types of manipulations, such as number of phonological features changed (e.g. White & Morgan, 2008), but our meta-analysis is underpowered to detect such effects.

Despite the proposal that infants should be more sensitive to consonant compared to vowel mispronunciations (Nazzi et al., 2016), we found no difference in sensitivity to consonant and vowel mispronunciations. But, a more nuanced picture was revealed regarding differences between consonant and vowel mispronunciations when further moderators were introduced. Sensitivity to consonant mispronunciations did not change with age and were similar for infants learning Germanic and Romance languages. In contrast, sensitivity to vowel mispronunciations increased with age and was greater overall for infants learning Germanic languages, although sensitivity to vowel mispronunciations did increase with age for infants learning Romance languages as well. These results show that sensitivity to vowel mispronunciations is modulated both by development and by native language, whereas sensitivity to consonant mispronunciations is fairly similar across age and native language.This pattern of results support previous experimental evidence that sensitivity to consonants and vowels have a different developmental trajectory and that this difference also depends on whether the infant is learning a Romance (French, Italian) or Germanic (British English, Danish) native language (Nazzi et al., 2016).

Our meta-analysis revealed the rather surprising result that onset overlap between labels for the target and distractor images lead to greater mispronunciation sensitivity in comparison to target-distractor pairs that shared no phonological overlap. It should be arguably more, not less, difficult to detect a mispronunciation (dag) when the target and distractor overlap in their onset phoneme (dog-duck), because the infant can not use differences in the onset sound between the target and distractor to identify the intended referent. Perhaps including overlap between the target and distractor lead infants to pay more attention to mispronunciations, leading to an increased effect of mispronunciation sensitivity. When we examined the distribution of this manipulation across infant age, however, we found an alternate explanation for this pattern of results. Older children were more likely to recieve the arguably more difficult manipulation where target-distractor pairs overlapped in their onset phoneme. If older children have greater mispronunciation sensitivity in general, then this may have lead to greater mispronunciation sensitivity for overlapping target-distractor pairs, instead of the manipulation itself.

But, our main developmental analysis found a lack of developmental change in mispronunciation sensitivity, suggesting that older children do not have greater mispronunciation sensitivity than younger children. If older children are given a more difficult task than younger children, however, this may dampen any developmental effects. It appears that this may be the case for overlap between target-distractor pairs. Older children were given a more difficult task (target-distractor pairs with onset overlap), which may have lowered the size of their mispronunciation sensitivity effect. Younger children were given an easier task (target-distractor pairs with no overlap), which may have relatively increased the size of their mispronunciation sensitivity effect. As a result, any developmental differences would be erased, hidden by task differences in the experiments that older and younger infants participated in. This argument is supported by the PRIMIR Framework (Curtin et al. (2011); Curtin and Werker (2007); Werker and Curtin (2005)), which argues that infants' ability to access the phonetic detail of familiar words is governed by the difficulty of their current task. Further support comes from evidence that sensitivity to mispronunciations when the target-distractor pair overlapped on the onset phoneme increased with age. This pattern of results suggest that when infants are given an equally difficult task, developmental effects may be revealed. This explanation can be confirmed by testing more young infants on overlapping target-distractor pairs.

## Data Analysis Choices

While creating the dataset on which this meta-analysis was based, we included as many details as possible to describe each study. During the coding of these characteristics, we noted a potential for variation in a handful of variables that relate to data analysis, specifically relating to timing (post-naming analysis window; onset time) and to the calculation of the dependent variable reported. We focused on these variables in particular because their choice can potentially be made after researchers have examined the data, leading to an inflated number of significant results which may also explain the publication bias observed in the funnel plot asymmetry analyses (Simmons et al., 2011). To explore whether this variation contributed to the lack of developmental change observed in the overall meta-analysis, we included these variables as moderators in a set of exploratory analyses. We noted an interesting pattern of results, specifically that different conclusions about mispronunciation sensitivity, but more notably mispronunciation sensitivity development, could be drawn depending on the length of the post-naming analysis window as well as the type of dependent variable calculated in the experiment (see Figures ?? and ??).

As infants' age increases, they recognize words more quickly (Fernald et al., 1998), which may lead experimenters to adjust and lower offset times in their analysis as well as shorten the length of the analysis window. Yet, we find no relationship between age and offset times, nor that offset time modulated mispronunciation sensitivity. Indeed, a majority of studies used an offset time between 360 and 370 ms, which follows the "best guess" of Swingley and Aslin (2000) for the amount of time needed for infants to initiate eye movements in response to a spoken target word. Without knowledge of the base reaction time in a given population of infants, use of this best guess reduces the number of free parameters used by researchers. In contrast, we found a negative correlation between infant age and the length of the post-naming analysis window, and that increasing the length of the post-naming analysis window decreases the size of mispronunciation sensitivity. Given a set of mispronunciation sensitivity data, a conclusion regarding the development of mispronunciation sensitivity would be different depending on the length of the post-naming analysis window. Although we have no direct evidence, an analysis window can be potentially set after collecting data. At worst, this adjustment could be the result of a desire to confirm a hypothesis, increasing the rate of false-positives (Gelman & Loken, 2013): a "significant effect" of mispronunciation sensitivity is found with an analysis window of 2000 but not 3000 ms, therefore 2000 ms is chosen. At best, this variation introduces noise into the study of mispronunciation sensitivity, bluring the true developmental trajectory of mispronunciation sensitivity. In the next section, we highlight some suggestions for how the field can remedy this issue.

Surpisingly, we found that the type of dependent variable calculated moderated mispronunciation sensitivity and conclusions about its developmental trajectory. Unlike the exploratory analyses related to timing, there is not a clear reason for one dependent variable to be chosen over another; the prevelence of each dependent variable appears distributed across ages and some authors always calculate the same dependent variable while others use them interchangeably in different publications. One clear difference is that both the Difference Score and Pre vs. Post dependent variables take into account each infants' actual preference in the pre-naming baseline phase, while the Post dependent variable does not. Without access to the raw data, it is difficult to conclusively determine why different dependent variable calculations influence mispronunciation sensitivity. In the next section, we advocate for the adoption of Open Data practices as one way to address this issue.

## Recommendations to Establish Analysis Standards

A lack of a field standard can have serious consequences, as our analyses show. On the one hand, this limits the conclusions we can draw regarding our key research question. Without access to the full datasets or analysis code of the studies included in this meta-analysis, it is difficult to pinpoint the exact role played by these experimental design and data analysis choices. On the other hand, this finding emphasizes that current practices of free, potentially ad hoc choices regarding data analyses are not sustainable if the field wants to move towards quantitative evidence for theories of language development.

We take this opportunity to suggest several recommendations to address the issue of potential posthoc analysis decisions. Preregistration can serve as proof of a priori decisions regarding data analysis, which can also contain a data-dependent description of how data analysis decisions will be made once data is collected. The peer-reviewed form of preregistration, termed Registered Reports, has already been adopted by a large number of developmental journals, and general journals that publish developmental works, showing the field's increasing acceptance of such practices for hypothesis-testing studies. Sharing data (Open Data) can allow others to re-analyze existing datasets to both examine the impact of analysis decisions and cumulatively analyze different datasets in the same way. Considering the specific issue of analysis time window, experimenters can opt to analyze the time course as a whole, instead of aggregating the proportion of target looking behavior. This allows for a more detailed assessment of infants' fixations over time and reduces the need to reduce the post-naming analysis window. Both Growth Curve Analysis (Law II & Edwards, 2015; Mirman, Dixon, & Magnuson, 2008) and Permutation Clusters Analysis (Delle Luche, Durrant, Poltrock, & Floccia, 2015; Maris & Oostenveld, 2007; Von Holzen & Mani, 2012) offer potential solutions to analyze the full time course. Furthermore, it may be useful to establish standard analysis pipelines for mispronunciation studies. This would allow for a more uniform analysis of this phenomenon, as well as aid experimenters in future research planning. In general, however, a better understanding of how different levels of linguistic knowledge may drive looking behavior is needed. We hope this understanding can be achieved by applying the above suggestions.

A meta-analysis is a first step in improving experiment planning by providing an estimate of the population effect and its variance, which is directly related to the sample needed to achieve satisfactory power in the null hypothesis significance testing framework. Failing to take effect sizes into account can lead to either underpowered research or testing too many participants. Underpowered studies will lead to false negatives more frequently than expected, which in turn results in an unpublished body of literature (Bergmann et al., 2018). At the same time, underpowered studies with significant outcomes are likely to overestimate the effect, leading to wrong estimations of the population effect when paired with publication bias (Jennions, Mù, Pierre, Curie, & Cedex, 2002). Overpowered studies mean that participants were tested unnecessarily, which has ethical implications particularly when working with infants and other difficult to recruit and test populations.

The estimated effect for mispronunciation sensitivity in this meta-analysis is 0.61, and the most frequently observed sample size is 24 participants. If we were to assume that researchers assess mispronunciation sensitivity in a simple paired t-test, the resulting power is 0.54. Reversely, to achieve 80% power, one would need to test 43.40 participants. These calculations suggest that for the comparison of responses for correct pronunciations and mispronunciations, the studies included in this meta-analysis are under-powered analyses. Furthermore, many studies in this meta-analysis included further factors to be tested, leading to two-way interactions (age versus mispronunciation sensitivity is a common example), which by some estimates require four times the sample size to detect an effect of similar magnitude as the main effect for both ANOVA (Fleiss, 1986) and mixed-effect-model (Leon & Heo, 2009) analyses. We thus strongly advocate for a consideration of power and the reported effect sizes to test infants' mispronunciation sensitivity.

## Conclusion

This meta-analysis comprises an aggregation of almost two decades of research on mispronunciation sensitivity, finding that infants accept both correct pronunciations and mispronunciations as labels for a target image. However, they are more likely to accept correct pronunciations, which indicates sensitivity to mispronunciations in familiar words. This sensitivity was not modulated by infant age or vocabulary, suggesting that from a young age on, before the vocabulary explosion, infants' word representations may be already phonologically well-specified. We recommend future theoretical frameworks take this evidence into account. Our meta-analysis was also able to confirm different findings in the literature, including the role of mispronunciation size, mispronunciation position, and the role of the native langauge in sensitivity to mispronunciation type (consonant vs. vowel). Furthermore, evidence of an interaction between task demands (phonological overlap between target-distractor pairs) and infant age may partially explain the lack of developmental change in mispronunciation sensitivity.

Despite this overall finding, however, we note evidence that data analysis choices can modulate conclusions about mispronunciation sensitivity development. Future studies should be carefully planned with this evidence in mind. Ideally, future experimental design and data analysis would become standardized which will be aided by the growing trend of preregistration and open science practices. Our analysis highlights how meta-analyses can aid in identification of issues in a particular field and play a vital role in how the field addresses such issues.

*Acknowledgements*: The authors would like to thank Emelyne Gaudichau for valuable assistance in entering data. Author 1 was supported by the Agence Nationale de la Recherche (ANR-13-BSH2-0004) and by training grant DC-00046 from the National Institute of Deafness and Communicative Disorders of the National Institutes of Health. Author 2 was supported by the European Horizon 2020 programme (Marie Skłodowska-Curie grant No 660911), the Agence Nationale de la Recherche (ANR-10-IDEX-0001-02 PSL*, ANR-10-LABX-0087 IEC) and the Fondation de France.

## References

Allaire, J., Xie, Y., McPherson, J., Luraschi, J., Ushey, K., Atkins, A., … Chang, W. (2018). rmarkdown: Dynamic Documents for R. Retrieved from https://cran.r-project.org/package=rmarkdown

Altvater-Mackensen, N. (2010). *Do manners matter? Asymmetries in the acquisition of manner of articulation features.* (PhD thesis). Radboud University Nijmegen.

Altvater-Mackensen, N., Feest, S. V. H. van der, & Fikkert, P. (2014). Asymmetries in early word recognition: The case of stops and fricatives. *Language Learning and Development*, *10*(2), 149–178. doi:10.1080/15475441.2013.808954

Aust, F., & Barth, M. (2018). papaja: Prepare reproducible APA journal articles with R Markdown. Retrieved from https://github.com/crsh/papaja

Bailey, T. M., & Plunkett, K. (2002). Phonological specificity in early words. *Cognitive Development*, *17*(2), 1265–1282. doi:10.1016/S0885-2014(02)00116-8

Ballem, K. D., & Plunkett, K. (2005). Phonological specificity in children at 1;2. *Journal of Child Language*, *32*(1), 159–173. doi:10.1017/S0305000904006567

Barton, D., Miller, R., & Macken, M. A. (1980). Do children treat clusters as one unit or two? In *Papers and reports on child language development* (pp. 93–137).

Bergelson, E., & Swingley, D. (2017). Young infants ' word comprehension given an unfamiliar talker or altered pronunciations. *Child Development*. doi:10.1111/cdev.12888

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*. doi:10.17605/OSF.IO/3UBNC

Bernier, D. E., & White, K. S. (2017). What's a Foo? Toddlers Are Not Tolerant of Other Children's Mispronunciations. In *Proceedings of the 41st annual boston university conference on language development* (pp. 88–100).

Bion, R. A. H., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30months. *Cognition*, *126*(1), 39–53. doi:10.1016/j.cognition.2012.08.008

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *Proceedings of the 39th annual conference of the cognitive science society* (pp. 124–129). Austin, TX: Cognitive Science Society, Inc. Retrieved from https://pdfs.semanticscholar.org/0807/41051b6e2b74d2a1fc2e568c3dd11224984b.pdf

Canfield, R. L., & Haith, M. M. (1991). Young infants' visual expectations for symmetric and asymmetric stimulus sequences. *Developmental Psychology*, *27*(2), 198–208. doi:10.1037/0012-1649.27.2.198

Charles-Luce, J., & Luce, P. A. (1995). An examination of similarity neighbourhoods in young children's receptive vocabularies. *Journal of Child Language*, *22*(3), 727–735. doi:10.1017/S0305000900010023

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences* (2nd ed.). New York: Lawrence Earlbaum Associates.

Csibra, G., Hernik, M., Mascaro, O., Tatone, D., & Lengyel, M. (2016). Statistical treatment of looking-time data. *Developmental Psychology*, *52*(4), 521–36. doi:10.1037/dev0000083

Curtin, S., Byers-Heinlein, K., & Werker, J. F. (2011). Bilingual beginnings as a lens for theory development: PRIMIR in focus. *Journal of Phonetics*, *39*(4), 492–504. doi:10.1016/j.wocn.2010.12.002

Curtin, S., & Werker, J. F. (2007). The perceptual foundations of phonological development. In M. G. Gaskell (Ed.), *The oxford handbook of psycholinguistics* (pp. 579–599). New York: Oxford University Press. doi:10.1093/oxfordhb/9780198568971.013.0035

Delle Luche, C., Durrant, S., Poltrock, S., & Floccia, C. (2015). A methodological investigation of the Intermodal Preferential Looking paradigm: Methods of analyses, picture selection and data rejection criteria. *Infant Behavior and Development*, *40*, 151–172. doi:10.1016/j.infbeh.2015.05.005

Feest, S. V. H. van der, & Fikkert, P. (2015). Building phonological lexical representations. *Phonology*, *32*(02), 207–239. doi:10.1017/S0952675715000135

Ferguson, C. J., & Heene, M. (2012). A vast graveyard of undead theories: Publication bias and psychological science's aversion to the null. *Perspectives on Psychological Science*, *7*(6), 555–561. doi:10.1177/1745691612459059

Fernald, A., Pinto, J. P., Swingley, D., Weinberg, A., & McRoberts, G. W. (1998). Rapid gains in speed of verbal processing by infants in the 2nd year. *Psychological Science*, *9*(3), 228–231. doi:10.1111/1467-9280.00044

Fernald, A., Swingley, D., & Pinto, J. P. (2001). When half a word is enough: infants can recognize spoken words using partial phonetic information. *Child Development*, *72*(4), 1003–15. doi:10.1111/1467-8624.00331

Fisher, R. A. (1922). On the Interpretation of $\chi$ 2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87. doi:10.2307/2340521

Fleiss, J. L. (1986). *The Design and Analysis of Clinical Experiments*. New York: Wiley; Sons.

Floccia, C., Nazzi, T., Luche, C. D., Poltrock, S., & Goslin, J. (2014). English-learning one- to two-year-olds do not show a consonant bias in word learning. *Journal of Child Language*, *41*(5), 1085–114. doi:10.1017/S0305000913000287

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. doi:10.1017/S0305000916000209

Gelman, A., & Loken, E. (2013). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time.* Department of Statistics, Columbia University. doi:10.1037/a0037714

Halberda, J. (2003). The development of a word-learning strategy. *Cognition*, *87*, B23–B34.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*(2), 107–128. doi:10.3102/10769986006002107

Højen, A., Madsen, T. O., Vach, W., Basbøll, H., Caporali, S., & Blese, D. (n.d.). *Contributions of vocalic and consonantal information when Danish 20-month-olds recognize familiar words.*

Jennions, M. D., Mù, A. P., Pierre, Â., Curie, M., & Cedex, F. P. (2002). Relationships fade with time : a meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society of London B: Biological Sciences*, *269*, 43–48. doi:10.1098/rspb.2001.1832

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. doi:10.1006/cogp.1995.1010

Keidel, J. L., Jenison, R. L., Kluender, K. R., & Seidenberg, M. S. (2007). Does grammar constrain statistical learning? *Psychological Science*, *18*(10), 922–923. doi:10.1111/j.1467-9280.2007.02001.x

Law II, F., & Edwards, J. R. (2015). Effects of Vocabulary Size on Online Lexical Processing by Preschoolers. *Language Learning and Development*, *11*(4), 331–355. doi:10.1080/15475441.2014.961066

Leon, A. C., & Heo, M. (2009). Sample sizes required to detect interactions between two binary fixed-effects in a mixed-effects linear regression model. *Computational Statistics and Data Analysis*, *53*(3), 603–608. doi:10.1016/j.csda.2008.06.010

Mani, N., Coleman, J., & Plunkett, K. (2008). Phonological specificity of vowel contrasts at 18-months. *Language and Speech*, *51*, 3–21. doi:10.1177/00238309080510010201

Mani, N., & Plunkett, K. (2007). Phonological specificity of vowels and consonants in early lexical representations. *Journal of Memory and Language*, *57*(2), 252–272. doi:10.1016/j.jml.2007.03.005

Mani, N., & Plunkett, K. (2010). Twelve-month-olds know their cups from their keps and tups. *Infancy*, *15*(5), 445–470. doi:10.1111/j.1532-7078.2009.00027.x

Mani, N., & Plunkett, K. (2011). Does size matter? Subsegmental cues to vowel mispronunciation detection. *Journal of Child Language*, *38*(03), 606–627. doi:10.1017/S0305000910000243

ManyBabiesConsortium. (n.d.). Quantifying sources of variability in infancy research using the infant-directed speech preference. *Advances in Methods and Practices in Psychological Science*.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG- and MEG-data. *Journal of Neuroscience Methods*, *164*(1), 177–190. doi:10.1016/j.jneumeth.2007.03.024

Markman, E. M., Wasow, J. L., & Hansen, M. B. (2003). Use of the mutual exclusivity assumption by young word learners. *Cognitive Psychology*, *47*(3), 241–275. doi:10.1016/S0010-0285(03)00034-3

Marslen-Wilson, W. D., & Zwitserlood, P. (1989). Accessing spoken words: The importance of word onsets. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 576–585. doi:10.1037/0096-1523.15.3.576

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*(1), 1–86. doi:10.1016/0010-0285(86)90015-0

Mills-Smith, L., Spangler, D. P., Panneton, R., & Fritz, M. S. (2015). A Missed Opportunity for Clarity: Problems in the Reporting of Effect Size Estimates in Infant Developmental Science. *Infancy*, *20*(4), 416–432. doi:10.1111/infa.12078

Mirman, D., Dixon, J. A., & Magnuson, J. S. (2008). Statistical and computational models of the visual world paradigm: Growth curves and individual differences. *Journal of Memory & Language*, *59*(4), 475–494. doi:10.1016/j.jml.2007.11.006

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., & Group, T. P. (2009). Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. *PLoS Medicine*, *6*(7), e1000097. doi:10.1371/journal.pmed.1000097

Morris, S. B., & DeShon, R. P. (2002). Combining effect size estimates in meta-analysis with repeated measures and independent-groups designs. *Psychological Methods*, *7*(1), 105–125. doi:10.1037/1082-989X.7.1.105

Nazzi, T., & Cutler, A. (2018). How Consonants and Vowels Shape Spoken-Language Recognition. *Annual Review of Linguistics*, (July), 1–23. doi:10.1146/annurev-linguistics

Nazzi, T., Floccia, C., Moquet, B., & Butler, J. (2009). Bias for consonantal information over vocalic information in 30-month-olds: Cross-linguistic evidence from French and English. *Journal of Experimental Child Psychology*, *102*(4), 522–537. doi:10.1016/j.jecp.2008.05.003

Nazzi, T., Poltrock, S., & Von Holzen, K. (2016). The developmental origins of the consonant bias in lexical processing. *Current Directions in Psychological Science*, *25*(4), 291–296. doi:10.1177/0963721416655786

Rabagliati, H., Ferguson, B., & Lew-Williams, C. (2018). The profile of abstract rule learning in infancy: Meta-analytic and experimental evidence. *Developmental Science*, (October 2017), 1–18. doi:10.1111/desc.12704

Ramon-Casas, M., & Bosch, L. (2010). Are non-cognate words phonologically better specified than cognates in the early lexicon of bilingual children? *Selected Proceedings of the 4th Conference on Laboratory Approaches to Spanish Phonology*, 31–36.

Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive Psychology*, *59*(1), 96–121. doi:10.1016/j.cogpsych.2009.02.002

R Core Team. (2018). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Renner, L. F. (2017). *The magic of matching – speech production and perception in language acquisition* (thesis). Stockholm University.

Sakaluk, J. (2016). Make it pretty: Forest and funnel plots for meta-analysis using ggplot2. [Blog post]. Retrieved from https://sakaluk.wordpress.com/2016/02/16/7-make-it-pretty-plots-for-meta-analysis/

Schwarzer, G. (2007). meta: An R package for meta-analysis. *R News*, *7*(3), 40–45. doi:10.1007/978-3-319-21416-0>

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

doi:10.1177/0956797611417632

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. doi:10.1037/a0033242

Skoruppa, K., Mani, N., Plunkett, K., Cabrol, D., & Peperkamp, S. (2013). Early word recognition in sentence context: French and English 24-month-olds' sensitivity to sentence-medial mispronunciations and assimilations. *Infancy*, *18*(6), 1007–1029. doi:10.1111/infa.12020

Stager, C. L., & Werker, J. F. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, *388*(6640), 381–382. doi:10.1038/41102

Swingley, D. (2009). Onsets and codas in 1.5-year-olds' word recognition. *Journal of Memory and Language*, *60*(2), 252–269. doi:10.1016/j.jml.2008.11.003

Swingley, D. (2016). Two-year-olds interpret novel phonological neighbors as familiar words. *Developmental Psychology*, *52*(7), 1011–1023. doi:10.1037/dev0000114

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166. doi:10.1016/S0010-0277(00)00081-0

Swingley, D., & Aslin, R. N. (2002). Lexical Neighborhoods and the Word-Form representations of 14-Month-Olds. *Psychological Science*, *13*(5), 480–484. doi:10.1111/1467-9280.00485

Tamasi, K. (2016). *Measuring children ' s sensitivity to phonological detail using eye tracking and pupillometry* (PhD thesis). University of Potsdam.

Tomasello, M., & Mervis, C. B. (1994). The instrument is great, but measuring comprehension is still a problem. In *Monographs of the society for research in child development* (pp. 174–179). doi:10.1111/j.1540-5834.1994.tb00186.x

Tsuji, S., Bergmann, C., & Cristia, A. (2014). Community-Augmented Meta-Analyses: Toward Cumulative Data Assessment. *Psychological Science*, *9*(6), 661–665. doi:10.1177/1745691614552498

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48. doi:10.18637/jss.v036.i03

Von Holzen, K., & Mani, N. (2012). Language nonselective lexical access in bilingual toddlers. *Journal of Experimental Child Psychology*, *113*, 569–586. doi:10.1016/j.jecp.2011.02.002

Werker, J. F., & Curtin, S. (2005). PRIMIR: A developmental framework of infant speech processing. *Language Learning and Development*, *1*(2), 197–234. doi:10.1207/s15473341lld0102_4

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, *52*(1), 114–132. doi:10.1016/j.jml.2008.03.001

Zesiger, P., Lozeron, E. D., Levy, A., & Frauenfelder, U. H. (2012). Phonological specificity in 12- and 17-month-old French-speaking infants. *Infancy*, *17*(6), 591–609. doi:10.1111/j.1532-7078.2011.00111.x

Table 1

*Summary of all studies. Age: truncation of mean age reported in the paper. Vocabulary: Comp = comprehension, Prod = production. Distractor Familiarity: Fam = Familiar Distractor, Unfam = Unfamiliar Distractor. Distractor Target Overlap: position of overlap between target and distractor; O = onset, M = medial, C = coda. Mispronunciation Size: number of features changed; commas indicate when sizes were compared separately (e.g. 1, 2, 3), dashes indicate the range of sizes were aggregated (e.g. 1-3). Mispronunciation Position: O = onset, M = medial, C = coda. Mispronunciation Type: C = consonant, V = vowel, T = tone. For both Mispronunciation Position and Type, a slash separator indicates that is was tested but a distinction was not made in the stimuli. For all categories, unspec. indicates that the value was unspecified in the paper*

| | | | | Distractor | | Mispronunciation | | | |
|---|---|---|---|---|---|---|---|---|---|
| Paper | Format | Age | Vocabulary | Familiarity | Target Overlap | Size | Position | Type | N Effect Sizes |
| Altvater-Mackensen (2010) | dissertation | 22, 25 | None | fam, unfam | O, novel | 1 | O, O/M | C | 13 |
| Altvater-Mackensen et al. (2014) | paper | 18, 25 | None | fam | O | 1 | O | C | 16 |
| Bailey & Plunkett (2002) | paper | 18, 24 | Comp | fam | none | 1, 2 | O | C | 12 |
| Bergelson & Swingley (2017) | paper | 7, 9, 12, 6 | None | fam | none | unspec | O/M | V | 9 |
| Bernier & White (2017) | proceedings | 21 | None | unfam | novel | 1, 2, 3 | O | C | 4 |
| Delle Luche et al. (2015) | paper | 20, 19 | None | fam | O | 1 | O | C/V | 4 |
| Durrant et al. (2014) | paper | 19, 20 | None | fam | O | 1 | O | C/V | 4 |
| HÃ¸jen et al. (n.d.) | gray paper | 19, 20 | Comp/Prod | fam | C, O | 2-3 | O/M, C/M | C/V, V, C | 6 |
| HÃ¶hle et al. (2006) | paper | 18 | None | fam | none | 1 | O | C | 4 |
| Mani & Plunkett (2007) | paper | 15, 18, 24, 14, 20 | Comp/Prod | fam | O | 1-2, 1 | O | V, C/V, C | 14 |
| Mani & Plunkett (2010) | paper | 12 | Comp | fam | O | 1 | M, O | V, C | 8 |
| Mani & Plunkett (2011) | paper | 23, 17 | None | unfam | novel | 1-3, 1, 2, 3 | M | V | 15 |
| Mani, Coleman, & Plunkett (2008) | paper | 18 | Comp/Prod | fam | O | 1 | M | V | 4 |
| Ramon-Casas & Bosch (2010) | paper | 24, 25 | None | fam | none | unspec | M | V | 4 |
| Ramon-Casas et al. (2009) | paper | 21, 20 | Prod | fam | none | unspec | M | V | 10 |
| Ren & Morgan (in press) | gray paper | 19 | None | unfam | none | 1 | O, C | C | 8 |
| Skoruppa et al. (2013) | paper | 23 | None | unfam | O/M | 1 | C | C | 4 |
| Swingley & Aslin (2000) | paper | 20 | Comp | fam | none | 1 | O | C/V | 2 |
| Swingley & Aslin (2002) | paper | 15 | Comp/Prod | fam | none | 1, 2 | O/M | C/V | 4 |
| Swingley (2003) | paper | 19 | Comp/Prod | fam | O | 1 | O, M | C | 6 |
| Swingley (2009) | paper | 17 | Comp/Prod | fam | none | 1 | O, C | C | 4 |
| Swingley (2016) | paper | 27, 28 | Prod | unfam | novel | 1 | O/M | C/V, C, V | 9 |
| Tamasi (2016) | dissertation | 30 | None | unfam | novel | 1, 2, 3 | O | C | 4 |
| Tao & Qinmei (2013) | paper | 12 | None | fam | none | unspec | unspec | T | 4 |
| Tao et al. (2012) | paper | 16 | Comp | fam | none | unspec | unspec | T | 6 |
| van der Feest & Fikkert, (2015) | paper | 24, 20 | None | fam | O | 1 | O | C | 16 |
| van der Feest & Johnson (2016) | paper | 24 | None | fam | O | 1 | O | C | 20 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Wewalaarachchi et al. (2017) | paper | 24 | None | unfam | novel | 1 | O/M/C | C/V/T, V, C, T | 8 |
| White & Aslin (2011) | paper | 18 | None | unfam | novel | 1 | M | V | 4 |
| White & Morgan (2008) | paper | 18, 19 | None | unfam | novel | 1, 2, 3 | O | C | 12 |
| Zesiger & JÃ¶hr (2011) | paper | 14 | None | fam | none | 1 | O, M | C, V | 7 |
| Zesiger et al. (2012) | paper | 12, 19 | Comp/Prod | fam | none | 1, 2 | O | C | 6 |

*Figure 1.* A PRISMA flowchart illustrating the selection procedure used to include studies in the current meta-analysis.

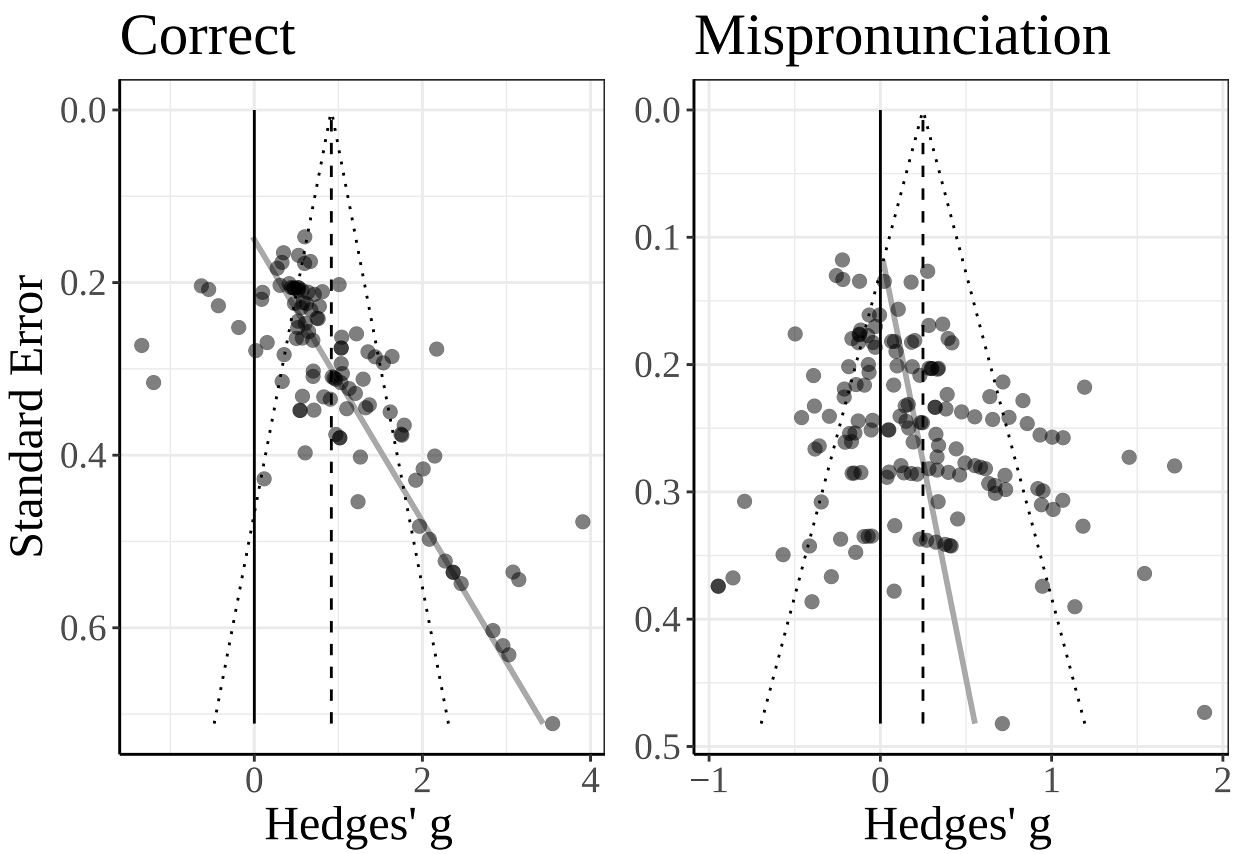


*Figure 2.* Funnel plots for object identification, plotting the standard error of the effect size in relation to the effect size. The black line marks zero, the dashed grey line marks the effect estimate, and the grey line marks funnel plot asymmetry.

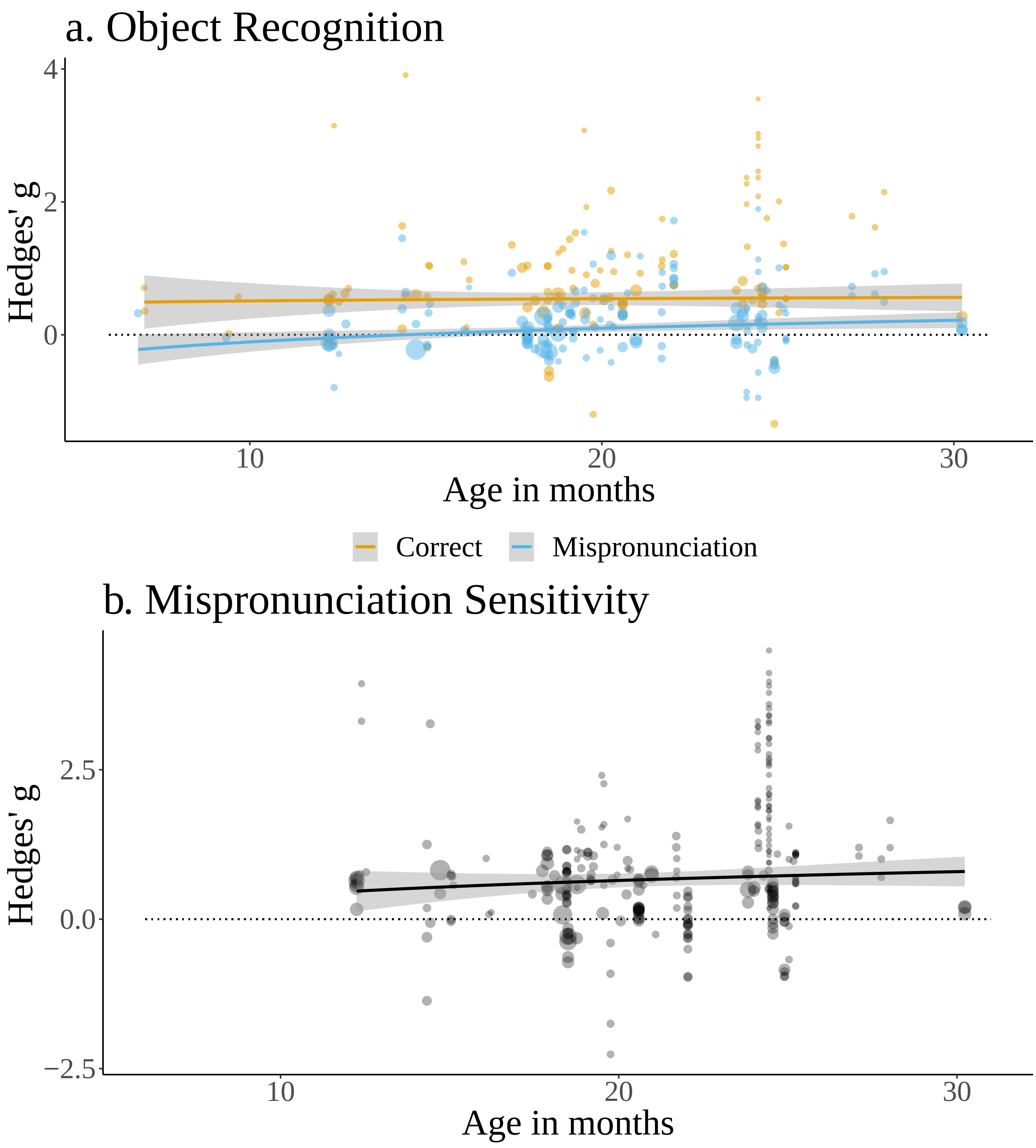


*Figure 3.* Panel a: Effect sizes for correct pronunciations (orange) and mispronunciations (blue) by participant age. Panel b: Effect sizes for mispronunciation sensitivity (correct - mispronunciations) by participant age. For both panels, point size depicts inverse variance and the dashed line indicates zero (chance).

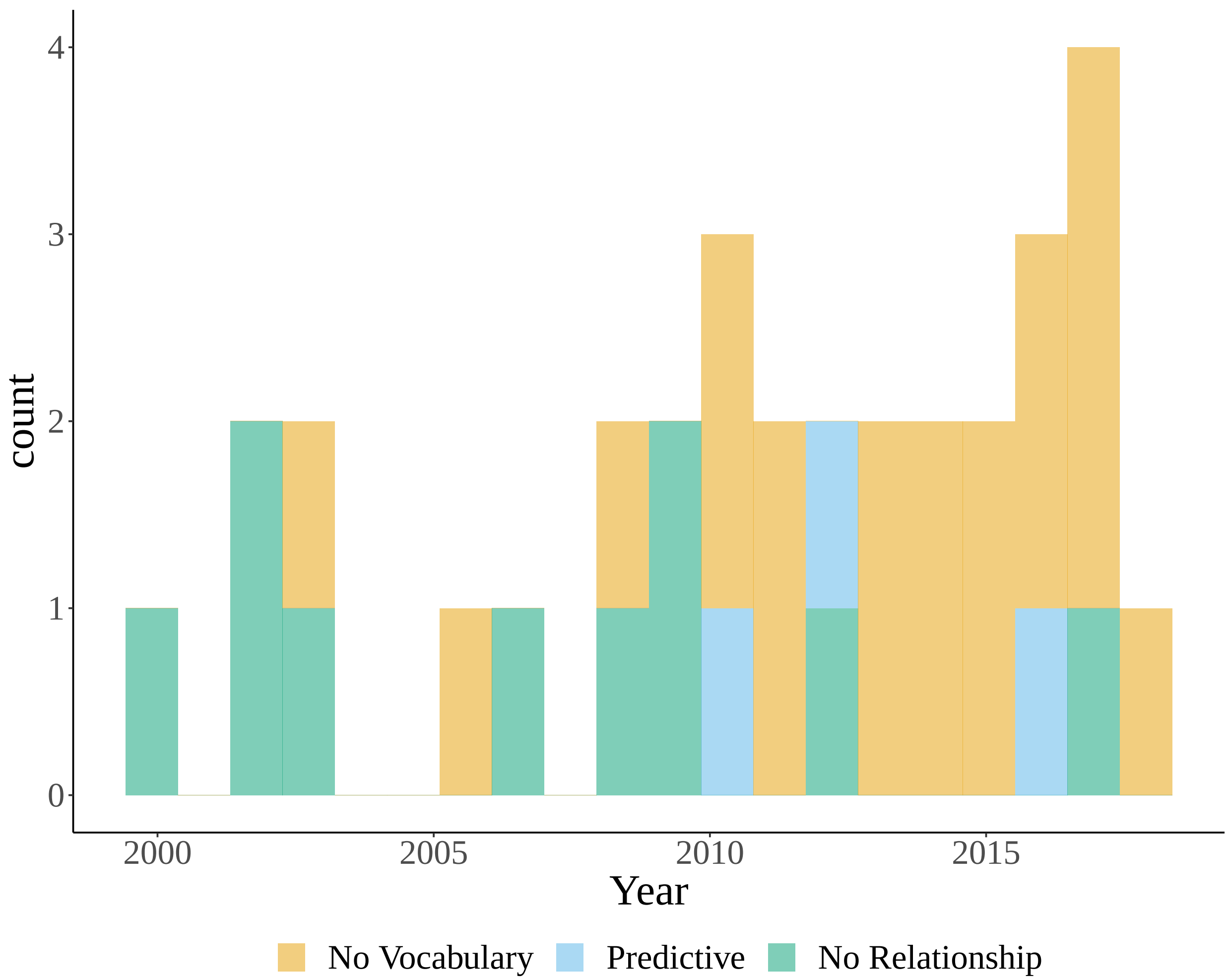


*Figure 4.* Counts of studies included in the meta-analysis as a function of publication year, representing whether the study did not measure vocabulary (orange), did measure vocabulary and was reported to predict mispronunciation sensitivity (blue), or did measure vocabulary and was reported to not predict mispronunciation sensitivity (green).

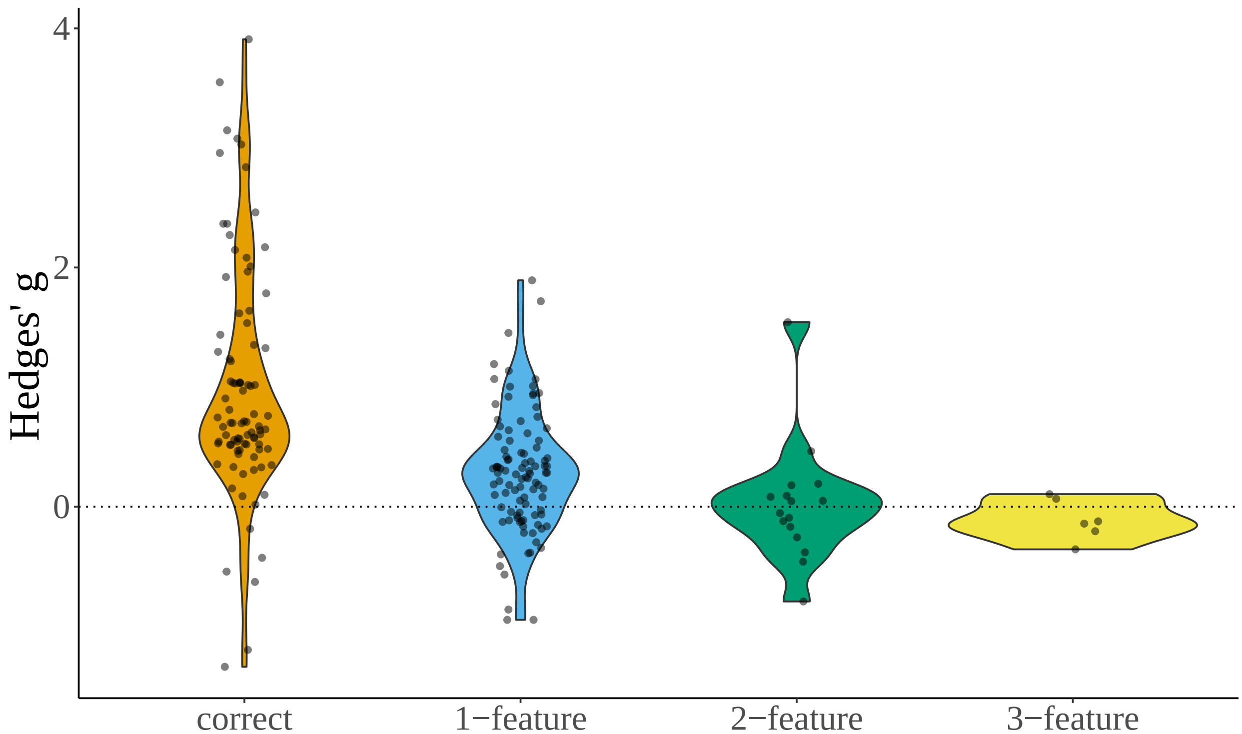


*Figure 5.* Effect sizes for correct pronunciations, 1-, 2-, and 3-feature mispronunciations.

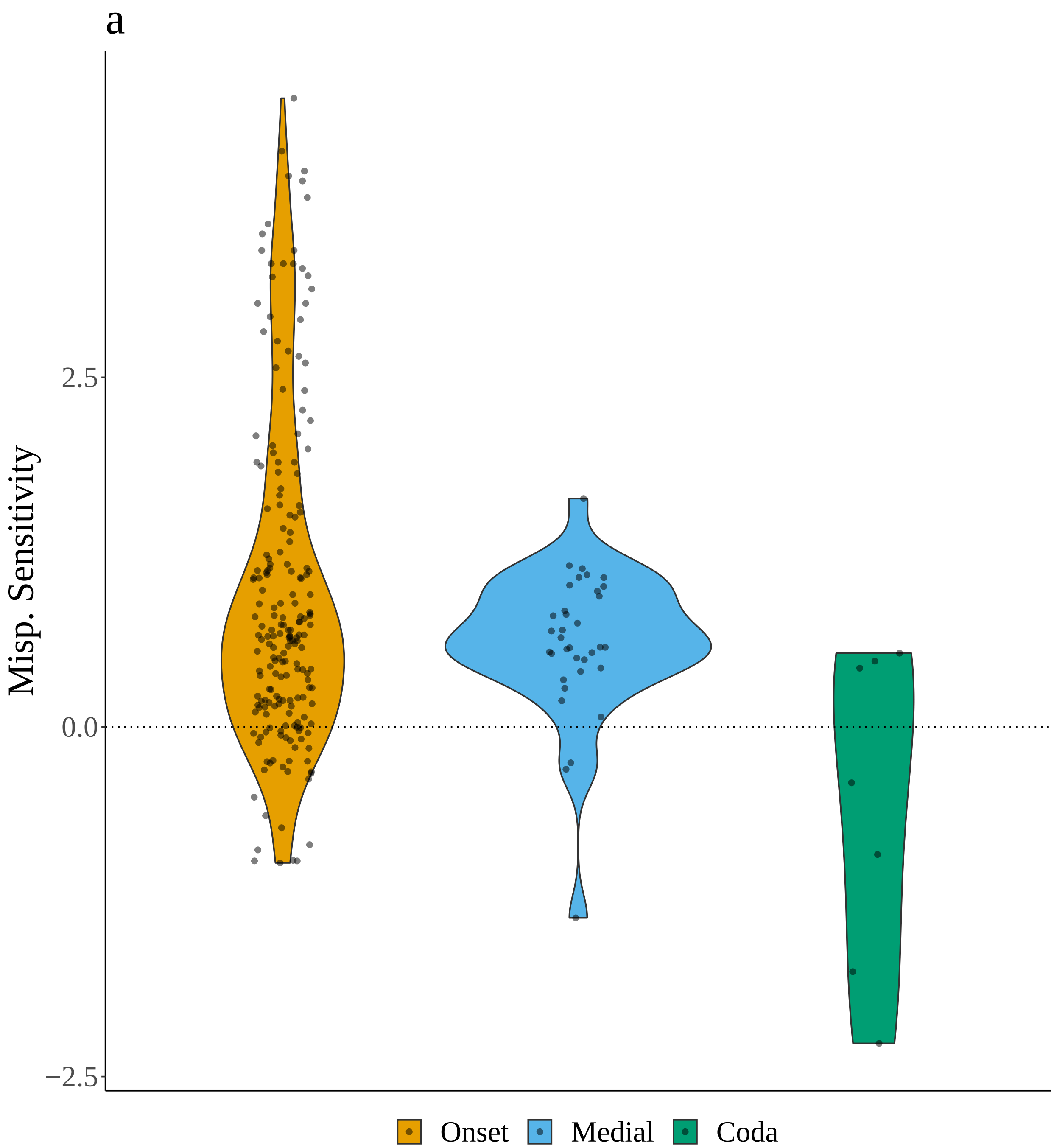


*Figure 6.* Panel a: Effect sizes for mispronunciation sensitivity (correct - mispronunciations) for mispronunciations on the onset, medial, and coda positions. Panel b: Effect sizes for mispronunciation sensitivity (correct - mispronunciations) for mispronunciations on the onset, medial, and coda positions by age. For both panels, point size depicts inverse variance and the dashed line indicates zero (chance).

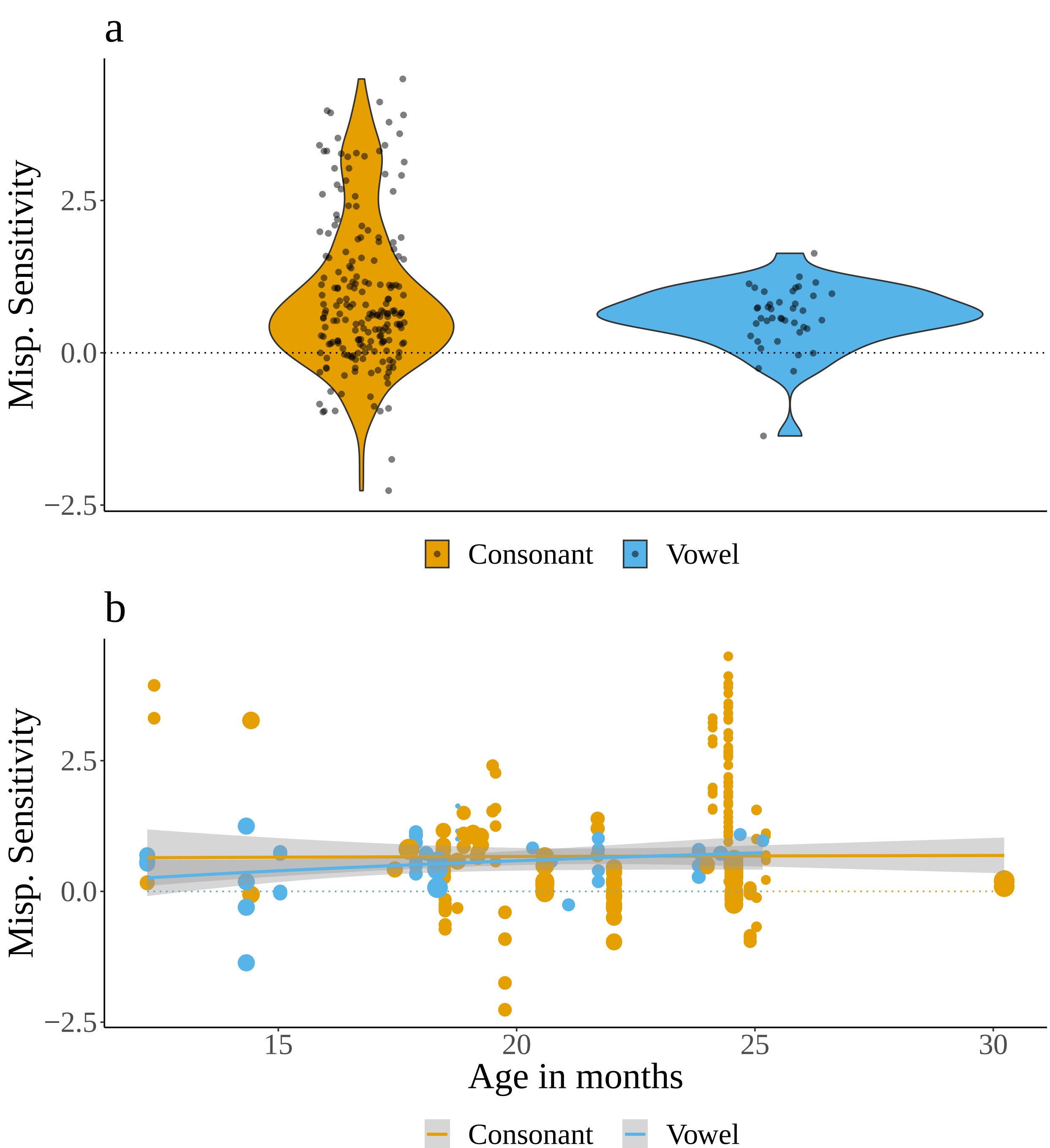


*Figure 7.* Panel a: Effect sizes for mispronunciation sensitivity (correct - mispronunciations) for consonant and vowel mispronunciations. Panel b: Effect sizes for mispronunciation sensitivity (correct - mispronunciations) for consonant and vowel mispronunciations by age. For both panels, point size depicts inverse variance and the dashed line indicates zero (chance).

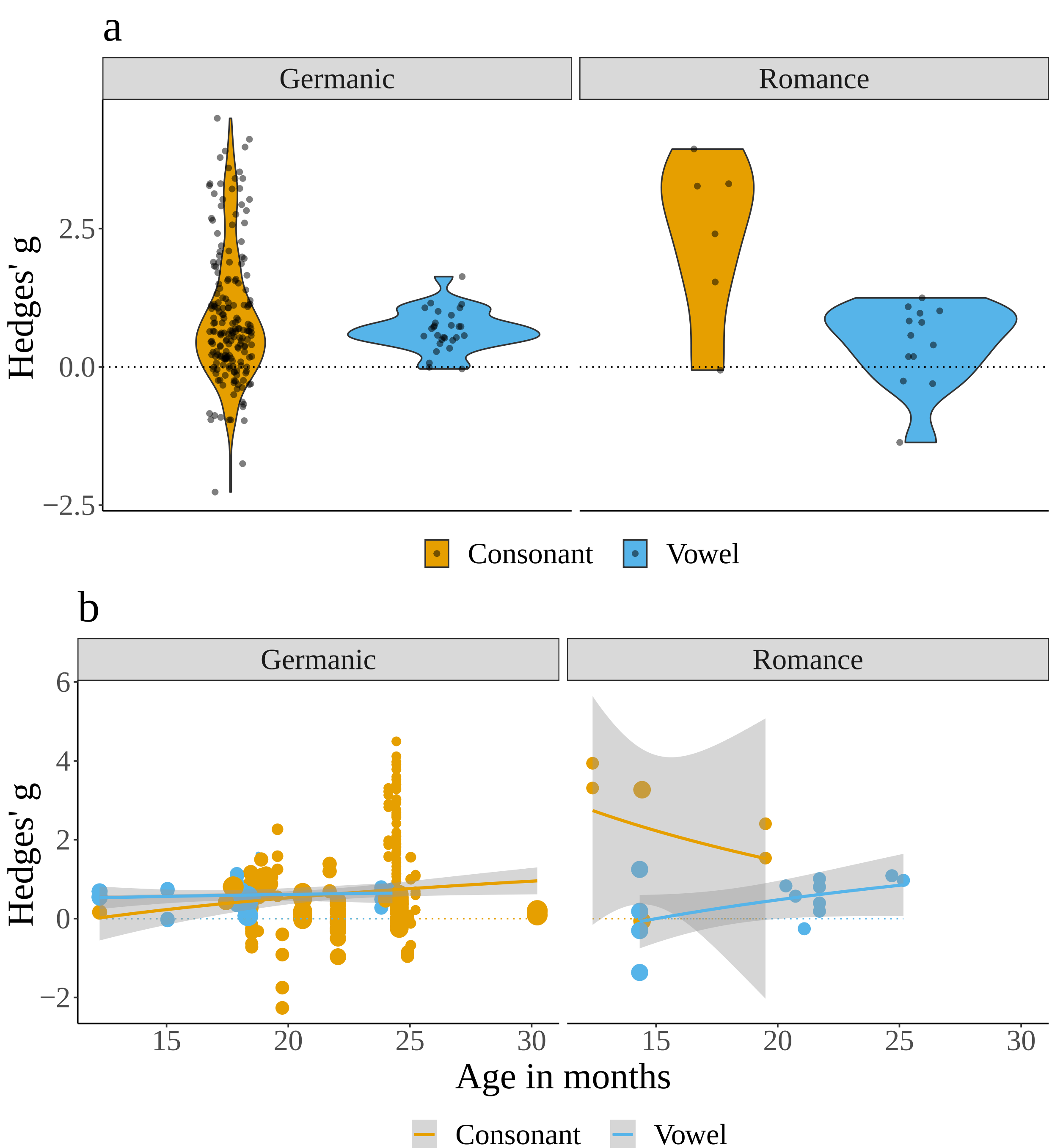


*Figure 8.* Panel a: Effect sizes for mispronunciation sensitivity (correct - mispronunciations) for consonant and vowel mispronunciations for infants learning a Germanic (left) or a Romance (right) native language. Panel b: Effect sizes for mispronunciation sensitivity (correct - mispronunciations) for consonant and vowel mispronunciations for infants learning a Germanic (left) or a Romance (right) native language by age. For both panels, point size depicts inverse variance and the dashed line indicates zero (chance).

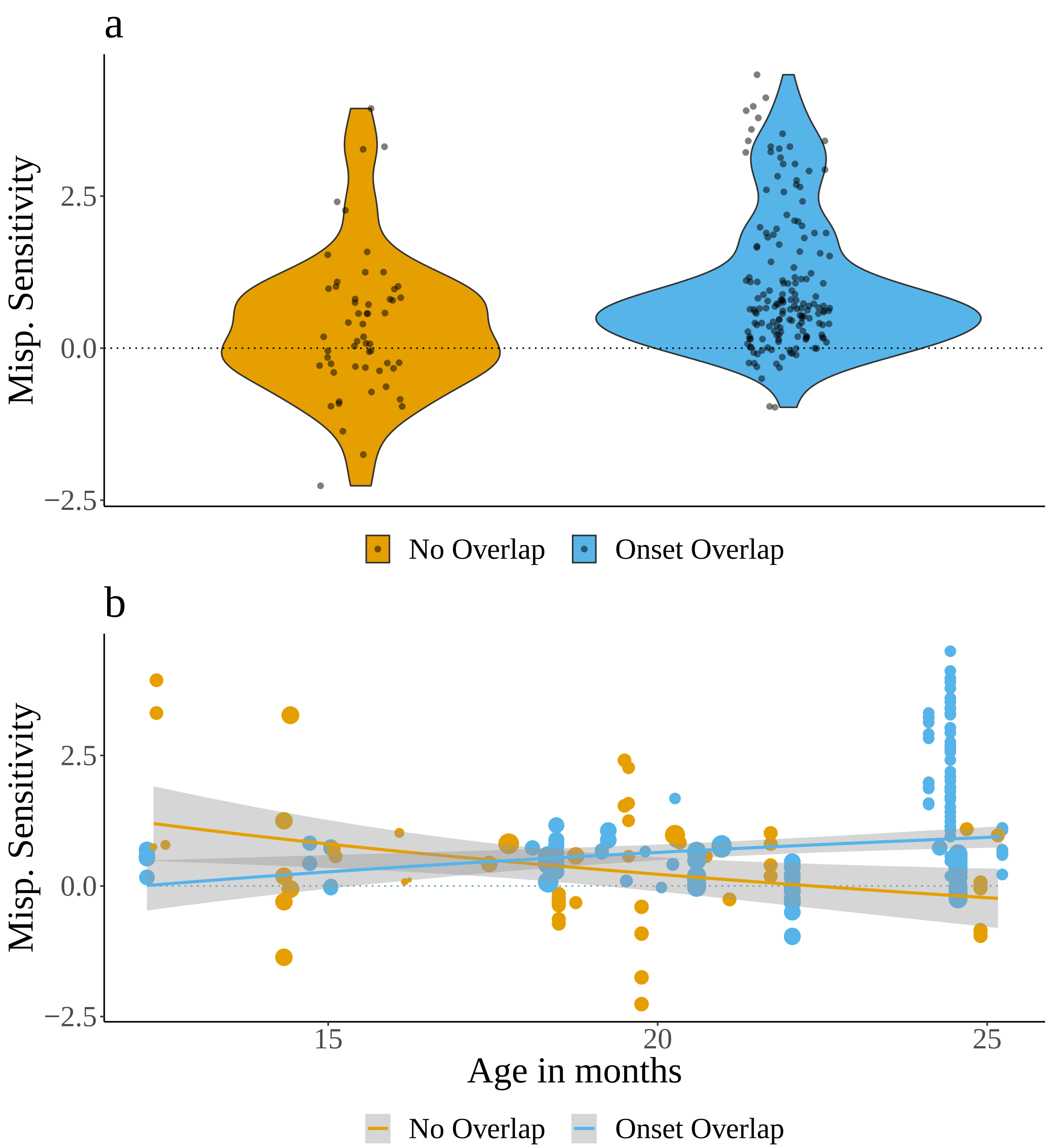


*Figure 9.* Panel a: Effect sizes for mispronunciation sensitivity (correct - mispronunciations) for target-distractor pairs with onset overlap or no overlap. Panel b: Effect sizes for mispronunciation sensitivity (correct - mispronunciations) for target-distractor pairs with onset overlap or no overlap by age. For both panels, point size depicts inverse variance and the dashed line indicates zero (chance).

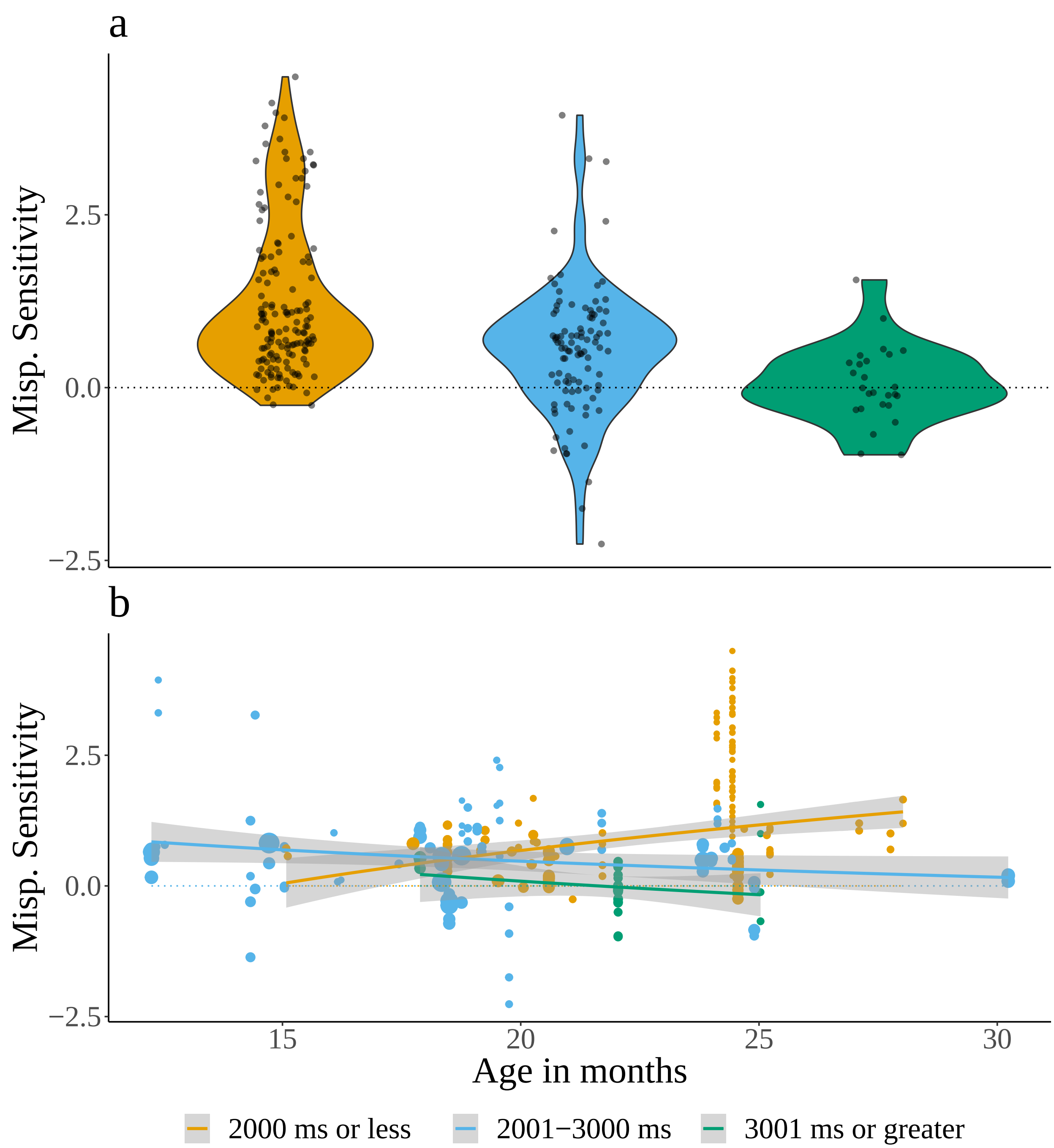


*Figure 10.* Effect sizes for the different lengths of the post-naming analysis window: 2000 ms or less (orange), 2001 to 3000 ms (blue), and 3001 ms or greater (green). Although length of the post-naming analysis window was included as a continuous variable in the meta-analytic model, it is divided into categories for ease of viewing. Panel a plots mispronunciation sensitivity aggregated over age, while panel b plots mispronunciation sensitivity as a function of age. The lines plot the linear regression and the gray shaded area indicates the standard error.

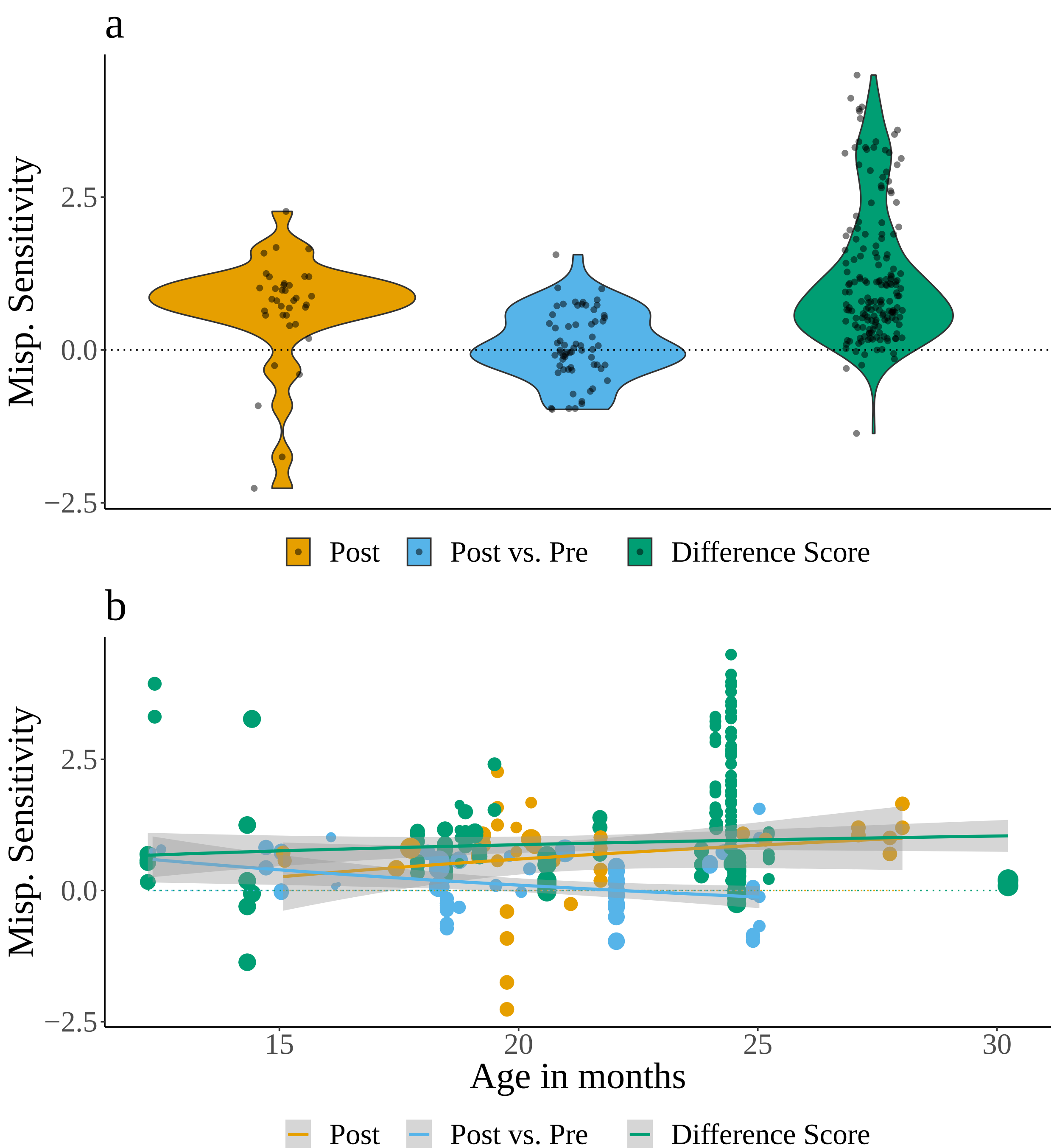


*Figure 11.* Effect sizes for the different types of dependent variables calculated: Post (orange), Post vs. Pre (blue), and Difference Score (green). Panel a plots mispronunciation sensitivity aggregated over age, while panel b plots mispronunciation sensitivity as a function of age. The lines plot the linear regression and the gray shaded area indicates the standard error.