Estimating age-related change in infants' linguistic and cognitive development using (meta-)meta-analysis

Anjie Cao[1], Molly Lewis[2], Sho Tsuji[3], Christina Bergmann[4], Alejandrina Cristia[3], & Michael C. Frank[1]

[1] Stanford University

[2] Carnegie Mellon University

[3] École Normale Supérieure - PSL

[4] Hochschule Osnabrück

Author Note

Enter author note here.

Correspondence concerning this article should be addressed to Anjie Cao, 450 Jane Stanford Way, Stanford, CA 94305. E-mail: anjiecao@stanford.edu

Abstract

Developmental psychology focuses on how psychological constructs change with age. In cognitive development research, however, the specifics of this emergence is often underspecified. Researchers often provisionally assume linear growth by including chronological age as a predictor in regression models. In this work, we aim to evaluate this assumption by examining the functional form of age trajectories across 24 phenomena in early linguistic and cognitive development using (meta-)meta-analysis, a statistical technique to combine the results of multiple meta-analyses. Surprisingly, for most meta-analyses, the effect size for the phenomenon did not change throughout development. We investigated four possible hypotheses explaining this pattern: (1) age-related selection bias against younger infants; (2) methodological adaptation for older infants; (3) change in only a subset of conditions; and (4) positive growth only after infancy. None of these explained the lack of age-related growth in most datasets. Our work challenges the assumption of linear growth in early cognitive development and suggests the importance of uniform measurement across children of different ages.

Keywords: keywords

Word count: X

Estimating age-related change in infants' linguistic and cognitive development using (meta-)meta-analysis

Developmental psychology focuses on how psychological constructs change with age. Throughout the years, many theories have been proposed to characterize and explain how and why developmental changes happen (Bronfenbrenner, 1977; Carey, 2009; Elman, 1996; Flavell, 1994; e.g., Piaget, 1971; Thelen & Smith, 2007). Among these theories, one common assumption is that skills refines with age (i.e. positive change assumption). Often, researchers treat age as a predictor in linear regression models, and therefore implicitly assume that the constructs of interests follow a linear trajectory (Lindenberger & Pötter, 1998).

One common approach to evaluating the functional form of age-related changes is through longitudinal observational studies. Measurements of psychological constructs, when tracked longitudinally, often reveal age trajectories that violate the linearity assumption. For instance, a longitudinal study that follows the development of executive function (EF) from 3 to 5 years-old using a battery of EF tasks show that EF follows a non-linear trajectory over age (Johansson, Marciszko, Brocki, & Bohlin, 2016). Similarly, vocabulary in early childhood, measured by MacArthur-Bates Communicative Development Inventories, also follows the exponential trend rather than the linear trend (Frank, Braginsky, Yurovsky, & Marchman, 2021). In many domains with established measurements, longitudinal observational research has been used to characterize the functional form of the development (Adolph, Robinson, Young, & Gill-Alvarez, 2008; Cole, Lougheed, Chow, & Ram, 2020; Karlberg, Engström, Karlberg, & Fryer, 1987; McArdle, Grimm, Hamagami, Bowles, & Meredith, 2009; Tilling, Macdonald-Wallis, Lawlor, Hughes, & Howe, 2014). However, longitudinal methods are more rarely applied to experimental studies that identify proposed mechanisms underlying development due to its resource intensiveness.

Many important findings in early language and cognitive development are primarily attested in cross-sectional experimental studies. For example, in the language learning domain, many studies have targeted specific mechanisms proposed to underlie how infants acquire specific facets of language. Constructs such as mutual exclusivity (Markman & Wachtel, 1988), statistical learning (Saffran, Aslin, & Newport, 1996), syntactic bootstrapping (Naigles, 1990) and so on, are all attested through decades of experimental evidence acquired through cross-sectional studies. These works are critical to test the causal mechanisms underlying age-related changes, but it is often challenging to make inferences about these changes from the observed effects.

First of all, the measurements properties of many experimental paradigms are rarely examined (Byers-Heinlein, Bergmann, & Savalei, 2022). When adapting an experimental design to test on a different age group, researchers often make adjustments based on intuitions or many trial-and-errors. For instance, an experimental paradigm that measures reaching might be too challenging for younger infants, so the researchers instead measures looking duration in the experiment that is designed to test on the same construct. If a looking time based paradigm elicit an effect in the piloting stage, then the looking time measurement would be adopted as the new measurement for the paradigm. If not, then more adaptations would follow. While these adjustments are sometimes fruitful, they might also inadvertently alter the psychometric properties of the experimental paradigm, changing the relationship between the latent constructs and the observed effects. Consequently, it would be difficult to compare infants' underlying abilities at one age – which are measured using an adapted paradigm – with infants' underlying abilities at another age.

Even when we assume the same psychometric properties of the experimental paradigms across age groups, many of the effects were rarely measured in samples with sufficient size and age variation to test the positive change assumption or the assumption of linearity in one individual study (cf. Frank et al., 2017). In an ideal world, one would run those experiments longitudinally on a large, diverse sample (Kidd & Garcia, 2022). In practice, this goal is difficult to achieve due to practical challenges, such as the constraints on both time and financial resources. As a result, the functional forms of age-related changes in critical constructs remain poorly understood.

To address this issue, we turned to meta-analysis. Meta-analysis is a statistical method to aggregate evidence across studies quantitatively. This approach has been widely adopted in many disciplines and subfields, including developmental psychology (Doebel & Zelazo, 2015; Hyde, 1984; Letourneau, Duffett-Leger, Levac, Watson, & Young-Morris, 2013). Compared to the single study approach, meta-analysis has several advantages. First, it allows us to examine the robustness of the phenomena documented in the literature. By combining results from multiple studies, meta-analysis enhances the statistical power to detect effects that might be too small to identify in individual studies. Second, meta-analysis provides a framework for assessing the consistency of research findings across different contexts (Borenstein, Hedges, Higgins, & Rothstein, 2021; Egger, Smith, & Phillips, 1997). Further, pooling across developmental studies with different cross-sectional samples may yield sufficient variation to explore the functional form of age-related change with greater precision than individual studies.

In this work, we aim to leverage meta-analysis to examine the shape of the developmental trajectory in key constructs in infant language and cognitive development. Specifically, we use existing meta-analyses from Metalab (https://langcog.github.io/metalab/), a platform that hosts community-augmented meta-analyses. Metalab was established to provide dynamic databases publicly available to all researchers (Bergmann et al., 2018). Researchers can deposit their meta-analysis dataset in the platform, and they can also use the dataset for custom analyses (Cao & Lewis, 2022; Cao, Lewis, & Frank, 2023; Lewis et al., 2016). To this date, Metalab contains 2967 effect sizes from 32 different meta-analysis and 48,529 unique participants, spanning different areas of developmental psychology [1]. This resource allows us to examine the suitability of meta-analysis as a tool to characterize developmental trajectory – and if suitable, provides insights into how these key constructs develop across the early months of childhood.

We acknowledge at the outset that meta-analysis has significant limitations. First of all, one significant issue in interpreting meta-analysis is the heterogeneity among studies. Heterogeneity refers to the variability in study participants, interventions, outcomes, and methodologies. This diversity, when originates from non-theoretically relevant variables, can make it challenging to aggregate results meaningfully, because differences between studies may reflect true variation in effects rather than a singular underlying effect size (Fletcher, 2007; Higgins & Thompson, 2002; Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006; Thompson & Sharp, 1999). Critically, understanding the source of heterogeneity often requires detailed coding of the potential moderators; this process is frequently hampered by the inadequate reporting standards prevalent in psychological literature, which often leaves essential information for coding these moderators absent (Nicholson, Deboeck, & Howard, 2017; Publications & Journal Article Reporting Standards., 2008). This process is also limited by the amount of studies available. If a moderator is only present in a few studies, then the lack of power would prohibit the testing of this moderator's influence. In addition, the quality of a meta-analysis is necessarily constrained by the quality of the existing studies (Eysenck, 1978; Simonsohn, Simmons, & Nelson, 2022). The effect sizes themselves can not inform us about the psychometric properties of the measurements, and the strengths of the effects might not truly reflect the strengths of the underlying constructs. If the studies being aggregated are flawed, the conclusions drawn from the meta-analysis will also be questionable. In summary, whether meta-analysis can provide insights into the nature of age-related change is dependent upon the quality and quantity of the existing literature.

This paper is organized as follows. In the first section, we provide an overview on the

[1] The snapshot of this dataset can be found in the github repository LINK

estimated general shape of age-related change across the datasets in Metalab. For each dataset, we compared the fit of the age model under four different functional forms: linear, logarithmic, quadratic, and constant. To preview our findings, we found that most datasets showed relatively constant effect size across age. For the datasets that showed a significant age effect, none of them supported the linearity assumption. In the second section, we tested four hypotheses on why many of the current meta-analyses failed to reveal age-related changes: (1) age-related selection bias against younger infants: more severe publication bias in studies testing younger infants, which results in more inflated effect sizes in younger infants; (2) methodological adaptation for older infants: making experiments more challenging for older infants, which results in diminished effect sizes in older infants; (3) change in only a subset of conditions: age effect is easier to detect in the subset of conditions with stronger effect sizes due to theoretical reasons; and (4) positive growth only after infancy. We found that none of the four explanations provided a satisfying account for the lack of age-related change in most meta-analyses consistently.

Estimating the functional forms of the developmental change in meta-analytic data

Datasets

Datasets were retrieved from Metalab. As of February 2024, Metalab hosted 32 datasets in total, with research areas ranging from language learning to cognitive development. Each dataset synthesized the literature in one research area, with the scope of the dataset determined by the original contributor of the dataset. All datasets included effect size estimates converted to Cohen's *d*, as well as estimates of effect size variance and a variety of other moderators (e.g., average age of participants) provided by the contributors. There were 2 desiderata for the datasets to be included in the final analysis:

1. The dataset must describe an experimental (non-correlational) effect that uses behavioral measures, and
2. For a dataset that has already been published, the aggregated meta-analytic effect reported in the published form must not be null (i.e., must be significantly different than zero).

Five datasets did not meet the first desideratum (*Pointing and vocabulary (concurrent)*; *Pointing and vocabulary (longitudinal)*; *Video deficit*; *Symbolic play*; *Word segmentation (neuro)*), and one dataset did not meet the second desideratum (*Phonotactic learning*). These datasets were not included in the analysis.

For the remaining 26 datasets, we made the following modifications to make their respective scope comparable such that each dataset corresponded to testing one distinct phenomenon. Following the organization in the original meta-analysis (Gasparini, Langus, Tsuji, & Boll-Avetisyan, 2021), we separated the *Language discrimination and preference* dataset into two datasets, one for discrimination and one for preference. We combined two pairs of datasets because they were testing the same experimental effects: *Gaze following* *(live)* and *Gaze following (video)* were combined into *Gaze following (combined)*; *Function word segmentation* and *Word segmentation (behavioral)* were combined into *Word segmentation (combined)*. We also replaced the *Infant directed speech preference* dataset with a more up-to-date version reported in Zettersten et al. (2024). Finally, for phenomenon that was predicted to follow a negative developmental trend (i.e. *Vowel discrimination (non-native)*), we flipped the sign of the effect sizes to make the effects comparable with the rest of the phenomena.

Our goal is to estimate the functional form of the developmental change in all of these meta-analytic datasets. To achieve this goal, we ran models with the same random effect structure specifications across all datasets. The random effect structure accounted for the both experiment-level grouping and the paper-level grouping. Not all datasets included these grouping variables so we recoded the missing ones to make sure the same random effect structure specifications could be applied to all datasets.

Since we were mostly interested in the age trajectory of these constructs in early childhood, we further trimmed the datasets to include only effect sizes from participants under 36 months of age. This decision did not qualitatively affect our findings as most datasets did not include data above age 36 months (94.33% of the effect sizes are from participants who were younger than 36 months of age). The final analysis included 25 datasets in total, each covers a different developmental time window. Table 1 presents the names of all the datasets, along with the number of effect sizes and participants included for each dataset.

Methods

All of the statistical analyses were conducted in R. Meta-analytic models were fit using the `metafor` package (Viechtbauer, 2010). This was an exploratory study in which no hypotheses were pre-registered. All of the analysis scripts and data are available at LINK.

For each dataset, we considered four functional forms as possible candidates for the shape of the developmental trajectory: linear, logarithmic, quadratic, and constant. A linear form is the most common assumption in the literature, whereas logarithmic and quadratic were chosen to represent sublinear growth and superlinear growth, respectively. The constant form served as a baseline for the other alternative growth patterns. Although other, more complex growth patterns are of course possible, we opted to compare these forms as a first pass. Note that the constant model includes one parameter (an intercept), linear and logarithmic models include two parameters (an intercept and a slope), and the quadratic model includes three parameters (intercept, slope, and quadratic growth term).

For all analyses, we fit multilevel random-effects meta-regression models using nested random intercepts to account for both the testing of the same infants in multiple conditions (e.g., in a between-participants design) and multiple studies within a single paper. Meta-regression models predicted effect sizes (Cohen's d) with mean age in months in different functional forms. We fit four meta-regression models in total for each dataset. These four models collectively test the positive change assumption and the linearity assumption. If the positive change assumption is true, we should expect the other three models all outperform the constant model. If the linearity assumption is true, the linear model would become the best fitting model for the dataset.

Results

Model comparison. Our initial goal was to compare the fit of models with different functional forms for each meta-analysis. Because models differed in their complexity (number of parameters), we extracted the corrected Akaike Information Criterion (AICc) for each model. AIC measures the quality of the model fits while penalizing models with more parameters. It is calculated as the difference between two times the number of estimated parameters in the model and two times the maximum value of the likelihood function of the model. The corrected AIC further adjusts for the number observations, which is particularly suitable when the sample size is small relative to the parameters in the model. The model with the lowest AICc was considered the best fitting model, and all the remaining models were compared against it. The remaining model each received a Δ_{AIC}, which was the difference between the AIC of the model and the AIC of the best fitting model. Following statistical convention, we treated $\Delta_{AIC} > 4$ as the statistical significance threshold (Burnham & Anderson, 2004). A best fitting model was significantly better than an alternative model if and only if the alternative model had $\Delta_{AIC} > 4$. Note that in the situation of a completely constant pattern of effects across age, the maximal difference in model fit would be an AICc of exactly 4 between the constant and quadratic model, because the former has one parameter and the latter has three parameters.

Figure 1 shows the prediction of each functional form. We found that the four functional forms could not be meaningfully distinguished in 19 out of 25 datasets. This situation typically arises because the data are constant and hence more complex models with zero parameters fit the data equally well. The remaining 6 datasets yielded meaningful contrasts between different functional forms, but the linear form was not the best-fitting form for any dataset. Table 2 shows the model comparison results for each dataset.

One limitation of the model comparison approach is that it does not quantify growth over time. To further examine the positive increase assumption, we estimated linear meta-regression models and examined the estimates on the age predictor. We found that the slope estimate for age was not significantly different from zero the in majority of the datasets (16/25; Figure 2).

Discussion

We conducted model comparisons to assess the functional forms of age-related change across 25 datasets. Four functional forms—linear, logarithmic, quadratic, and constant—were largely indistinguishable within most datasets. Notably, in the 6 datasets where evidence for a better fit of growth models compared to the constant model were found, linear models received no support, challenging the prevalent linearity assumption for early linguistic and cognitive development.

Further, in direct statistical assessment of positive increases over age using regression models, we only detected evidence for linear growth in 9/25 meta-analyses. Past work has successfully revealed age-related changes using meta-analysis (e.g. Best & Charness, 2015; McCartney, Harris, & Bernieri, 1990; Sugden & Marquis, 2017). But in most datasets that we have considered, effect size does not increase with age, despite theoretical reason to assume such a developmental change over age in the phenomena considered. Why?

Here we consider four explanations for the lack of age-related change in most of the meta-analyses we examined. First, meta-analyses are susceptible to publication bias, thus a tendency for studies showing effects or larger effects in the expected direction to be preferably published (Ferguson & Brannick, 2012; Ferguson & Heene, 2012; Francis, 2012; Mathur & VanderWeele, 2021; Thornton & Lee, 2000). And the bias could be related to the characteristics of the study, such as the inclusion of younger participants (Kathleen M. Coburn & Vevea, 2015). Researchers might have stronger incentives to publish positive results from younger infants since these results are sometimes perceived as more novel. Consequently, studies with younger participants may have effect sizes that were more inflated, compared to the studies with older participants. The selectivity of publication bias would thus obscure the possible developmental changes in the dataset (Figure 3, Panel 1).

Second, researchers may adapt methods as infants get older. Older infants have larger behavioral repertories, can stay attentive for longer period of time, and are in general better learners. As a results, studies that test older infants might have more demanding designs (Figure 3, Panel 2). For example, the high-amplitude sucking paradigm is most likely to be deployed on very young infants, whereas the paradigm measuring infants' looking time is most likely to be used on older infants. We did see some evidence for method adaptation in some datasets. For example, in *Language discrimination*, the average age for studies using a sucking paradigm (e.g. Christophe & Morton, 1998) was 0.58 months (SD = 0.89), but 5.30 months (SD = 1.78) for studies using looking time paradigm (e.g. Chong, Vicenik, & Sundara, 2018). This age-related change in research paradigms could lead to a case of Simpson's paradox: the age-related trend within a single method might be lost when multiple methods are combined (Kievit, Frankenhuis, Waldorp, & Borsboom, 2013; Simpson, 1951).

Third, other theoretical factors unrelated to age could also contribute to the lack of developmental effects. Some meta-analyses specifically tested whether these factors would moderate the effect sizes. For instance, in *Syntactic Bootstrapping*, the effect was only present in studies with transitive conditions (Cao & Lewis, 2022), In *Familiar word recognition*, the effect was stronger in infants whose primary language exposure was from Romance languages (Carbajal, Peperkamp, & Tsuji, 2021). Perhaps the apparent lack of developmental effects in the current analysis could be attributed to theoretical reasons, rather than a true absence of developmental changes (Figure 3, Panel 3). We were able to investigate these potential moderating effects in 22 of 25 datasets since these datasets were published manuscripts.

Fourth, developmental change in infancy and early childhood might be distinct from one another. Bergelson (2020) has speculated that word comprehension in the looking-while-listening paradigm only shows significant developmental changes after 12 months of age, with infants younger than 12 months showing mostly flat developmental trajectories in this task. This contrast could be attributed to the fact that older infants are not only more experienced compared to younger infants, but also better learners who can more effectively take advantage of the input they receive. There is much evidence suggesting that developmental changes occurring in one domain would have cumulative, cascading effects on changes in other domains (Ahmed, Kuhfeld, Watts, Davis-Kean, &

Vandell, 2021; Bornstein, Hahn, Putnick, & Pearson, 2018; Oakes & Rakison, 2019). The outcome of such developmental cascades might not be measurable in the experimental tasks included in the meta-analyses until infants are above 12 months of age (Figure 3, Panel 4).

We investigate each of these explanations in turn, assessing empirical support in our data. We summarise the results of these analyses in Table 3; in brief, no explanation provided traction for more than a small number of datasets.

Understanding the lack of developmental change in meta-analytic data

Age-related selection bias against younger infants

We first consider whether age-related selection bias can explain the lack of developmental changes in our datasets. If studies with younger infants suffered from publication bias more, then their effect sizes would be more inflated, obscuring possible developmental changes.

Methods. There are many methods to detect publication bias. One of the most common approaches is Egger's test (Egger, Smith, Schneider, & Minder, 1997), which examines the relationship between the studies' effect sizes and their precision. A significant result from Egger's test indicated an asymmetry in the funnel plot, suggesting the presence of publication bias. This method is more sensitive than the rank correlation approach, another common publication bias detection method (Begg & Mazumdar, 1994). However, Egger's test cannot accommodate predictors other than the study's precision. As a result, we also turned to the weight-function model developed by Vevea and Hedges (1995). This method detects publication bias by likelihood ratio tests: a bias-corrected model is pitted against the original model to see if the former provides a better fit than the latter. A positive result indicates the presence of publication bias.

To detect age-related publication bias, we splitted each dataset by the median of the average participant age associated with each effect size (in months). Median was chosen because age in the datasets was not normally distributed. We then run both Egger's test and the weight-function model on each half of the dataset. We compared the test outcomes from both tests across the two halves of the datasets. For Egger's test, we used the `regtest` function implemented in `metafor` (Viechtbauer, 2010). For the weight-function model, we used the package `weightr` (Kathleen M. Coburn & Vevea, 2019) and specified random-effect meta-regression models predicting effect sizes with mean age in months.

Egger's test was run on all but the 4 datasets in which either half of the datasets contained less than 20 effect sizes. Previous study has shown that Egger's test has reduced sensitivity in datasets with less than 20 studies (Sterne, Egger, & Smith, 2001). For similar reasons, 7 datasets were excluded in the weight-function analysis.

Results and discussion. Egger's test suggested that in 3 datasets there was evidence for publication bias in the younger half but not in the older half (*Audio-Visual Congruence*, *Categorization bias*, *Syntactic bootstrapping*). However, this result was not corroborated by the weight-function analysis. For these three datasets, the weight function analysis did not find evidence for publication bias in either half of the three datasets. This suggests that the significant results found by Egger's test might be due to factors other than publication bias. The weight-function analysis only found evidence for publication bias in the younger half but not the older half in one dataset: *Language Preference* (Younger: $\chi^2 = 6.08$, $p = 0.01$; Older: $\chi^2 = 3.27$, $p = 0.07$). This dataset yielded no significant results for either half in Egger's test.

We also further explored whether splitting dataset by 12-month of age would yield different patterns. In this follow-up analysis, we did not find any evidence for age-related publication bias using Egger's test. When using the weight-function analysis, only *Prosocial agent* showed publication bias in younger infants but not older infants (Younger: $\chi^2 = 8.02$, $p < 0.01$; Older: $\chi^2 = 0.30$, $p = 0.58$)

Overall, we found little evidence for more severe publication bias among the younger infants. The Egger's test and the function-weight analysis did not yield converging evidence, suggesting that factors other than publication bias may be at play in contributing to the results.

Methodological adaptation for older infants

In experiments with young children, many design decisions are made to ensure the paradigms are age appropriate (Byers-Heinlein et al., 2022). For older children, more behavioral measures are available and longer experiments are made possible by increased attention span. As a result, experimenters might test more subtle experimental contrasts. Perhaps the increasing difficulty or subtlety of experimental conditions for older infants mask age-related increase in effect sizes related to a particular construct. For example, imagine that different experimenters wanted to study word learning with 12- and 24-month-olds. The experimenter working with the younger group might choose a paradigm in which only two novel words were taught, while the experimenter working with the older children might choose to teach four. The resulting effect for older children might be weaker despite overall improvement in the underlying construct.

The accessibility of different methods could also potentially cause an instance of Simpson's paradox (Kievit et al., 2013). Imagine there were two methods, method A and method B, with the former having lower task demands than the latter. Due to its low task demands, method A would be more likely to be used on younger infants and causes larger effect sizes. In contrast, method B would be more likely to be used on older infants and results in smaller effect sizes. Although the age trend could be positive within each method, when pooling across studies from the two methods, the trend would then be negative, canceling out age-related changes patterns.

Since it is difficult to code for task demands across all studies, we explored whether methodological adaptation influences the developmental trend from the other side: instead of looking at method adaptation with age, we focused on studies using identical methods to test multiple age groups (e.g., Tsuji & Cristia, 2014). This subset of data should provide the best chance of detecting age-related changes in the absence of methodological variation.

Methods

We first needed to identify the subset of studies in each dataset that satisfy the following two criteria: (1) the same paper tested multiple age groups, and (2) the multiple age groups were all tested using the same experimental design and measure. The first criterion was operationalized as having a paper with multiple age groups with an age difference greater than one month. The second criterion was operationalized based on methodological moderators coded by the original authors and available in MetaLab.

Within the effects selected for each dataset, we calculated Δ_{age} for each effect size. Δ_{age} was the difference between the age associated with a particular effect size and the minimum age in each subset of the dataset.

19 datasets had subsets of studies fitting our criteria. We focused on the 15 subsets that had 10 and more effect sizes. For each subset, we applied a multilevel meta-regression model using the same nested random intercept as previously described. The model predicts effect sizes based on Δ_{age}. This analysis follows the logic that, if on average there is a greater effect size when the same experiment is conducted with older children relative to younger children, then the relation of effect size to Δ_{age} should be positive. Note that here we were only testing the linear assumption since it was the most parsimonious assumption.

Results and discussion

We found no significant relationship between Δ_{age} and the effect sizes in any of the dataset (all $p > 0.05$). In addition, we also explored whether there is a relationship between age and effect sizes in these datasets. In *Statistical sound category learning*, *Online* *word recognition* and *Mutual exclusivity*, the relationship between age and effect sizes was significant in the subset currently explored. However, these datasets also contained significant age effect in the full dataset. In other words, subsetting the datasets into containing only studies that tested multiple age groups using the same experimental paradigm did not reveal more age-related trends.

This analysis was necessarily constrained by the granularities of the coded moderators. The number of coded methodological moderators ranged from 1 to 9, which means that the experimental design was reduced into at maximum 9 dimensions. However, even at 9 dimensions, it is possible that elements of experiment design influencing task demands were overlooked. For instance, in many domains that use visual stimuli, the particular choice of visual stimuli might significantly vary in complexity (e.g. Cao & Lewis, 2022). Visual complexity has long been proposed as a key factor influencing the task demands (Hunter & Ames, 1988; Kosie et al., 2023), but stimulus complexity was not coded in any of our meta-analyses. In conclusion, the findings presented here should be interpreted with caution due to potential limitations in the coding of methodological moderators.

Change in only a subsest of conditions

Across the 25 datasets, 22 datasets were published through manuscripts in peer-reviewed venues. Among these manuscripts, we found that 8 papers reported that the meta-analytic effect was significantly stronger in a subset of the data. The subset was often identified by a particular condition in the experimental paradigm (e.g. experiment that shows “giving and taking action” to infants, Margoni & Surian, 2018), or certain characteristics of the participants (e.g. bilingual infants, Tsui, Byers-Heinlein, & Fennell, 2019). In the rest of the data, the meta-analytic effect was either significantly weaker or not present at all. There are many reasons for why the effect would be stronger or only present in a subset of the data. Here, we remained agnostic to the underlying causes for these differences, and leveraged these findings to ask: Is it possible that the influence of age was only observable in the subset of the dataset characterized by stronger effect sizes? Perhaps noise in other conditions inadvertently masked age-related changes.

Methods. We screened through 22 papers and identified 8 papers that reported a stronger effect on subsets of the data. All subsets had more than 10 effect sizes. For datasets reporting more than one subset as having stronger effect, we consider each respectively. In sum, 7 datasets produced 9 subsets that showed stronger effects.

We first investigated whether we could confirm the original patterns, i.e. the effect sizes in the better halves were indeed stronger than the other halves. To this end, we splitted the meta-analyses, where one subset was claimed to show the expected effect, and the other consisted of the remainder of the data (n $>$ 10 across all subsets). We ran the same multilevel meta-regression without any predictor to estimate the meta-analytic effect sizes in each half. Then we ran a Wald test to compare the two estimates by running a fixed-effects meta-regression model predicting effect sizes with the moderator distinguishing the two halves. A significant estimate on the moderator indicates that the meta-analytic effect sizes in both halves are significantly different from one another. We then estimated the slope of the age predictor in a multilevel meta-regression model for each of the subsets with larger effect sizes.

Results and discussion. We did not fully confirm the effect reported in the original papers: the "better half" identified by the original meta-analysis did not produce significantly stronger effects than the rest of the data in 7 datasets. We did observe a significantly stronger effect in the remaining 3 datasets: For *Prosocial Agents*, there was a stronger effect in experimental paradigms showing infants giving-taking actions compared to the studies showing infants other stimuli (Margoni & Surian, 2018, z = -2.47, p = 0.01); For *Statistical Sound Category Learning*, stronger effect was observed in studies using habituation paradigm compared to other paradigms (Cristia, 2018, z = -2.42, p = 0.02), and for *Statistical word segmentation*, stronger effect was observed in studies labeled as the conceptual replication of the original work (Black & Bergmann, 2017, $z = 2.51$, $p = 0.01$).

In addition, we did not find constraining our analyses to the "better half" increased the number of significant slope estimates. The two significant slope estimates came from *Mutual Exclusivity* ($\beta = 0.04$, $SE = 0.01$, $z = 4.63$, $p < 0.01$) and *Statistical sound category learning* ($\beta = 0.11$, $SE = 0.05$, $z = 2.23$, $p = 0.03$), which also showed significant slopes in the analyses with the full datasets. Qualitatively, we did see that the estimates increased in magnitude in *Syntactic bootstrapping* ($\beta = 0.01$, $p = 0.67$) and *Switch task* ($\beta = 0.01$, $p = 0.79$), but neither reached the statistical significance threshold.

The discrepancy between our analyses and the previously reported findings could be attributed to the different statistical models we chose – in the original meta-analysis papers, the models tend to differ in their particular specification of the nested random effect structure and in the inclusions of moderators. We chose the simplest model with the maximum random effect structure per recommendation (Barr, Levy, Scheepers, & Tily, 2013). This approach ensured fair comparison across all datasets, but it could diminish the strength of the reported effects.

Interestingly, even in the datasets where the better half effect was confirmed, we failed to see a significant age effect in datasets that did not show age-related changes in the original full dataset (*Prosocial agents* and *Statistical word segmentation*). Altogether, this set of analysis suggested that the theoretical constraints on the effect sizes could not adequately explain the lack of age-related change.

Positive growth only after infancy

Last but not least, we considered whether there is evidence for discontinuity between the growth patterns in infancy and beyond. Bergelson (2020)'s hypothesis on the development of word comprehension suggests a notable shift post the 12-month mark in infancy. This raises the question of whether such distinctions extend across various tasks.

This section aims to delve into these dynamics by only looking at the subset of the dataset with infants older than 12-month-olds.

Methods. Similar to previous analyses, we filtered each dataset to include only studies that reported more than 10 effect sizes that tested infants older than 12 months. 15 datasets met the criteria and contained sufficient effect sizes from participants above 12 months of age. We ran the same meta-regressions predicting effect size with mean age in months on this subset, and then we compared the estimates on the age predictor with the same models run on the full datasets.

Results and discussion. If the discontinuity account is true, we should expect to see more significant age effects to emerge on models run on the subset of data with older infants. We found support for this hypothesis in two datasets, *Cross-situational Word Learning* ($\beta = 0.01$, $SE < 0.01$, $z = 2.71$, $p = 0.01$) and *Mispronunciation sensitivity* ($\beta = 0.07$, $SE = 0.01$, $z = 4.69$, $p < 0.01$). In both datasets, there were no age effects in the full datasets, but significant age-related change in the subsets with older infants. This suggests that the discontinuity hypothesis was supported in at least two datasets. However, it is also worth noting that we also found the opposite patterns. In *Categorization bias* and *Sound symbolism*, there was evidence for age-related change across the entire age range, but no evidence for age-related change in the toddler subset (Both $p > 0.05$).

General discussion

How do infants' cognitive and linguistic abilities change with age? In this work, we leveraged a dataset of meta-analyses to evaluate the assumption that these abilities increase positively with age, and that the form of this increase is linear. There was no evidence for linear growth in 16 datasets, and interestingly, in all of these datasets, there was no evidence for any age-related growth at all. For the rest of the 9 datasets, none of them had the linear model as the best-fitting model.

In the second section, we investigated four potential explanations for this pattern: (1) age-related selection bias against younger infants; (2) methodological adaptation for older infants; (3) change in only a subset of conditions; and (4) positive growth only after infancy. We showed that none of these hypotheses provide explanations for the lack of age-related growth in most datasets. Table 3 shows a summary of whether each hypothesis can explain the lack of linear growth in each dataset.

Our current work has several strengths. By leveraging a large dataset of meta-analyses, we were able to conduct a comprehensive investigation of the developmental trend in cognitive and linguistic development across a wide range of domains. This broad, data-driven approach allows us to identify potential common patterns that might be missed in smaller studies. It also reveals that the linear form is not the best functional form to describe the developmental trajectories in datasets that showed a significant age related change. Furthermore, our investigation of the four hypotheses provides a thorough exploration of potential factors influencing the lack of developmental trend we observed. These analyses ensure that our conclusions are well-supported by the data.

At the same time, our current work has several limitations. First and foremost, we simply lacked sufficient data to investigate the possible explanations for many domains (see Table 3). In many datasets, when we filtered datasets to answer the corresponding questions, we lacked sufficient data to adequately test our hypotheses. As with many meta-analyses, our datasets also had high residual heterogeneity, meaning that we can only explain relatively small amounts of the variation among effect sizes, even when taking theoretically relevant factors into account.

Our work highlights the importance of improving reporting standards in developmental psychology. Testing moderation of heterogeneity requires consistent coding of moderators across datasets. But surveys of reporting standards show that many potential moderators go unreported. For instance, fewer than half of papers report attrition rate (Nicholson et al., 2017; Raad, Bellinger, McCormick, Roberts, & Steele, 2007). Given these observations, there is a clear need for the developmental psychology community to create and embrace more rigorous and transparent reporting standards. Researchers could follow the open science practices and make their stimuli more publicly available. This would enable other researchers to conduct follow-up analysis such as investigating the visual complexity of the stimuli. In addition, the recently developed framework for reporting demographics information across cultures in developmental psychology is also one promising direction moving forwards (Singh et al., 2023). Learning from other fields could provide valuable insights into how to enhance these standards. In biomedical research, numerous reporting standards have been published and widely adopted (for clinical trials: CONSORT, Schulz, Altman, & Moher, 2010; for epidemiological research: STROBE, Vandenbroucke et al., 2007; for meta-analysis and systematic review, PRISMA: Moher et al., 2015; for a catalog of reporting guidelines in health research: EQUATOR, Altman, Simera, Hoey, Moher, & Schulz, 2008). Following these structured guidelines in reporting could significantly increase both the quality and the quantity of information extractable from the original papers, providing more traction for tackling heterogeneity in meta-analysis.

Our work also underscores the importance of multi-laboratory large scale replication projects. The relationship between meta-analysis and multi-laboratory is complicated (Kvarven, Strømland, & Johannesson, 2020; Lewis, Mathur, VanderWeele, & Frank, 2022). Although the latter approach is much more time- and resource- intensive than the former, it is also much more effective in controlling unwanted heterogeneity and detecting subtle patterns in the data. One prominent example is the comparison between the meta-analysis of Infant directed speech preference (Dunst, Gorman, & Hamby, 2012) and the ManyBabies 1 project on the same topic (Consortium, 2020). Zettersten et al. (2024) found that, after an update to the meta-analysis dataset, both datasets yielded comparable estimated effect sizes ($d = 0.35$), but the age effect was only detected in the ManyBabies 1 project, not the meta-analysis. The study speculated that our second explanation (methodological variation covarying with age) might account for their results. In our analysis, we did investigate the methodological adaptation hypothesis in the IDS preference dataset. However, the methodological moderators available for us were limited and we could not incorporate the varying nature of the stimuli into our analysis. This example shows the potential limitations of meta-analyses that rely on aggregated data from studies with varied methodologies. In contrast, multi-laboratory collaboration projects like ManyBabies (Visser et al., 2022) can rely on standardized data collection procedure and stimuli, therefore providing a more controlled dataset to answer a specific research question with high power.

It is also worth considering whether the strengths of certain developmental phenomena truly stay constant throughout the first years of life. First of all, this counterintuitive possibility casts doubts on the construct validity of the existing measures. Many researchers strive to build on existing experimental procedures and measurements when they are testing participants of different age. This then leads to a potentially problematic situation: an experimental paradigm could have high construct validity with participants of a certain age, but low construct validity with participants of different age (e.g. Rovee-Collier & Cuevas, 2009). This is consistent with our analysis showing no evidence for developmental change even within studies using the same methods as indicated by the coded methodological moderators. As a result, a conundrum emerges: methodological adaptation could be a source of significant heterogeneity, obscuring the measurable developmental change. But at the same time, paradoxically, it could also be the prerequisite for properly measuring developmental change. This dilemma calls attention to the importance of properly examining the psychometric properties of the measures used in cross-sectional developmental psychology research.

Last but not least, an alternative explanation for the lack of developmental change is the limited sensitivity of the cross-sectional design. The group average may stay constant, but there could still be growth in an individual's performance across development

(Bornstein, Putnick, & Esposito, 2017). The nuanced nature of developmental change might be best captured by dense, longitudinal data of individual child (e.g. Bergelson et al., 2023; Sullivan, Mei, Perfors, Wojcik, & Frank, 2021).

In sum, our current work presents a surprising finding concerning age-related change in the cognitive and language development literature in early childhood. Despite decades of research built upon the positive increase and linearity assumptions, we failed to find evidence supporting either in most meta-analyses that we had access to. Our work is not intended to overturn the longstanding developmental theories. Like other researchers, we believe that infants get better across different cognitive and linguistic domains as they get older. Instead, our work aims to highlight the needs for more robust reporting standards and more large-scale multi-laboratory projects that measure children consistently across age groups and over time. Our findings invite the cognitive development community to strengthen our understanding of foundational assumptions via collaborative efforts.

References

Adolph, K. E., Robinson, S. R., Young, J. W., & Gill-Alvarez, F. (2008). What is the shape of developmental change? *Psychological Review*, *115*(3), 527.

Ahmed, S. F., Kuhfeld, M., Watts, T. W., Davis-Kean, P. E., & Vandell, D. L. (2021). Preschool executive function and adult outcomes: A developmental cascade model. *Developmental Psychology*, *57*(12), 2234.

Altman, D. G., Simera, I., Hoey, J., Moher, D., & Schulz, K. (2008). EQUATOR: Reporting guidelines for health research. *The Lancet*, *371*(9619), 1149–1150.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3), 255–278.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, 1088–1101.

Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are better learners. *Child Development Perspectives*, *14*(3), 142–149.

Bergelson, E., Soderstrom, M., Schwarz, I.-C., Rowland, C. F., Ramirez-Esparza, N., Hamrick, L. R., et al.others. (2023). *Everyday language input and production in 1001 children from 6 continents*.

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

Best, R., & Charness, N. (2015). Age differences in the effect of framing on risky choice: A meta-analysis. *Psychology and Aging*, *30*(3), 688.

Black, A., & Bergmann, C. (2017). Quantifying infants' statistical word segmentation: A meta-analysis. *39th Annual Meeting of the Cognitive Science Society*, 124–129. Cognitive Science Society.

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Bornstein, M. H., Hahn, C.-S., Putnick, D. L., & Pearson, R. M. (2018). Stability of core language skill from infancy to adolescence in typical and atypical development. *Science Advances*, *4*(11), eaat7422.

Bornstein, M. H., Putnick, D. L., & Esposito, G. (2017). Continuity and stability in development. *Child Development Perspectives*, *11*(2), 113–119.

Bronfenbrenner, U. (1977). Toward an experimental ecology of human development. *American Psychologist*, *32*(7), 513.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods & Research*, *33*(2), 261–304.

Byers-Heinlein, K., Bergmann, C., & Savalei, V. (2022). Six solutions for more reliable infant research. *Infant and Child Development*, *31*(5), e2296.

Cao, A., & Lewis, M. (2022). Quantifying the syntactic bootstrapping effect in verb learning: A meta-analytic synthesis. *Developmental Science*, *25*(2), e13176.

Cao, A., Lewis, M., & Frank, M. C. (2023). A synthesis of early cognitive and language development using (meta-) meta-analysis. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *45*.

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

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

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

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

Coburn, Kathleen M., & Vevea, J. L. (2015). Publication bias as a function of study characteristics. *Psychological Methods*, *20*(3), 310.

Coburn, Kathleen M., & Vevea, J. L. (2019). *Weightr: Estimating weight-function models for publication bias.*

Cole, P. M., Lougheed, J. P., Chow, S.-M., & Ram, N. (2020). Development of emotion regulation dynamics across early childhood: A multiple time-scale approach. *Affective Science*, *1*, 28–41.

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

Cristia, A. (2018). Can infants learn phonology in the lab? A meta-analytic answer. *Cognition*, *170*, 312–327.

Doebel, S., & Zelazo, P. D. (2015). A meta-analysis of the dimensional change card sort: Implications for developmental theories and the measurement of executive function in children. *Developmental Review*, *38*, 241–268.

Dunst, C., Gorman, E., & Hamby, D. (2012). Preference for infant-directed speech in preverbal young children. *Center for Early Literacy Learning*, *5*(1), 1–13.

Egger, M., Smith, G. D., & Phillips, A. N. (1997). Meta-analysis: Principles and procedures. *Bmj*, *315*(7121), 1533–1537.

Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *Bmj*, *315*(7109), 629–634.

Elman, J. L. (1996). *Rethinking innateness: A connectionist perspective on development* (Vol. 10). MIT press.

Eysenck, H. J. (1978). *An exercise in mega-silliness.*

Ferguson, C. J., & Brannick, M. T. (2012). Publication bias in psychological science: Prevalence, methods for identifying and controlling, and implications for the use of meta-analyses. *Psychological Methods*, *17*(1), 120.

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.

Flavell, J. H. (1994). *Cognitive development: Past, present, and future.*

Fletcher, J. (2007). What is heterogeneity and is it important? *Bmj*, *334*(7584), 94–96.

Francis, G. (2012). Publication bias and the failure of replication in experimental psychology. *Psychonomic Bulletin & Review*, *19*, 975–991.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., et al.others. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy*, *22*(4), 421–435.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project.* MIT Press.

Gasparini, L., Langus, A., Tsuji, S., & Boll-Avetisyan, N. (2021). Quantifying the role of rhythm in infants' language discrimination abilities: A meta-analysis. *Cognition*, *213*, 104757.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558.

Huedo-Medina, T. B., Sánchez-Meca, J., Marín-Martínez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q statistic or i^2 index? *Psychological Methods*, *11*(2), 193.

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

Hyde, J. S. (1984). How large are gender differences in aggression? A developmental meta-analysis. *Developmental Psychology*, *20*(4), 722.

Johansson, M., Marciszko, C., Brocki, K., & Bohlin, G. (2016). Individual differences in early executive functions: A longitudinal study from 12 to 36 months. *Infant and Child Development*, *25*(6), 533–549.

Karlberg, J., Engström, I., Karlberg, P., & Fryer, J. G. (1987). Analysis of linear growth using a mathematical model: I. From birth to three years. *Acta Paediatrica*, *76*(3), 478–488.

Kidd, E., & Garcia, R. (2022). How diverse is child language acquisition research? *First Language*, *42*(6), 703–735.

Kievit, R., Frankenhuis, W. E., Waldorp, L., & Borsboom, D. (2013). Simpson's paradox in psychological science: A practical guide. *Frontiers in Psychology*, *4*, 54928.

Kosie, J., Zettersten, M., Abu-Zhaya, R., Amso, D., Babineau, M., Baumgartne, H., et al.others. (2023). *ManyBabies 5: A large-scale investigation of the proposed shift from familiarity preference to novelty preference in infant looking time*.

Kvarven, A., Strømland, E., & Johannesson, M. (2020). Comparing meta-analyses and preregistered multiple-laboratory replication projects. *Nature Human Behaviour*, *4*(4), 423–434.

Letourneau, N. L., Duffett-Leger, L., Levac, L., Watson, B., & Young-Morris, C. (2013). Socioeconomic status and child development: A meta-analysis. *Journal of Emotional and Behavioral Disorders*, *21*(3), 211–224.

Lewis, M., Braginsky, M., Tsuji, S., Bergmann, C., Piccinini, P. E., Cristia, A., et al. (2016). *A quantitative synthesis of early language acquisition using meta-analysis*.

Lewis, M., Mathur, M. B., VanderWeele, T. J., & Frank, M. C. (2022). The puzzling relationship between multi-laboratory replications and meta-analyses of the published literature. *Royal Society Open Science*, *9*(2), 211499.

Lindenberger, U., & Pötter, U. (1998). The complex nature of unique and shared effects in hierarchical linear regression: Implications for developmental psychology. *Psychological Methods*, *3*(2), 218.

Margoni, F., & Surian, L. (2018). Infants' evaluation of prosocial and antisocial agents: A meta-analysis. *Developmental Psychology*, *54*(8), 1445.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

Mathur, M. B., & VanderWeele, T. J. (2021). Estimating publication bias in meta-analyses of peer-reviewed studies: A meta-meta-analysis across disciplines and journal tiers. *Research Synthesis Methods*, *12*(2), 176–191.

McArdle, J. J., Grimm, K. J., Hamagami, F., Bowles, R. P., & Meredith, W. (2009). Modeling life-span growth curves of cognition using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods*, *14*(2), 126.

McCartney, K., Harris, M. J., & Bernieri, F. (1990). Growing up and growing apart: A developmental meta-analysis of twin studies. *Psychological Bulletin*, *107*(2), 226.

Moher, D., Shamseer, L., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., … Group, P.-P. (2015). Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-p) 2015 statement. *Systematic Reviews*, *4*, 1–9.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*(2), 357–374.

Nicholson, J. S., Deboeck, P. R., & Howard, W. (2017). Attrition in developmental psychology: A review of modern missing data reporting and practices. *International Journal of Behavioral Development*, *41*(1), 143–153.

Oakes, L. M., & Rakison, D. H. (2019). *Developmental cascades: Building the infant mind.* Oxford University Press.

Piaget, J. (1971). *The theory of stages in cognitive development.*

Publications, A., & Journal Article Reporting Standards., C. B. W. G. on. (2008). Reporting standards for research in psychology: Why do we need them? What might they be? *The American Psychologist*, *63*(9), 839.

Raad, J. M., Bellinger, S., McCormick, E., Roberts, M. C., & Steele, R. G. (2007). Brief report: Reporting practices of methodological information in four journals of pediatric and child psychology. *Journal of Pediatric Psychology*, *33*(7), 688–693.

Rovee-Collier, C., & Cuevas, K. (2009). Multiple memory systems are unnecessary to account for infant memory development: An ecological model. *Developmental Psychology*, *45*(1), 160.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Schulz, K. F., Altman, D. G., & Moher, D. (2010). CONSORT 2010 statement: Updated guidelines for reporting parallel group randomised trials. *Journal of Pharmacology and Pharmacotherapeutics*, *1*(2), 100–107.

Simonsohn, U., Simmons, J., & Nelson, L. D. (2022). Above averaging in literature reviews. *Nature Reviews Psychology*, *1*(10), 551–552.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *13*(2), 238–241.

Singh, L., Barokova, M. D., Baumgartner, H. A., Lopera-Perez, D. C., Omane, P. O., Sheskin, M., et al.others. (2023). A unified approach to demographic data collection for research with young children across diverse cultures. *Developmental Psychology*.

Sterne, J. A., Egger, M., & Smith, G. D. (2001). Investigating and dealing with publication and other biases in meta-analysis. *Bmj*, *323*(7304), 101–105.

Sugden, N. A., & Marquis, A. R. (2017). Meta-analytic review of the development of face discrimination in infancy: Face race, face gender, infant age, and methodology moderate face discrimination. *Psychological Bulletin*, *143*(11), 1201.

Sullivan, J., Mei, M., Perfors, A., Wojcik, E., & Frank, M. C. (2021). SAYCam: A large, longitudinal audiovisual dataset recorded from the infant's perspective. *Open Mind*, *5*, 20–29.

Thelen, E., & Smith, L. B. (2007). Dynamic systems theories. *Handbook of Child Psychology*, *1*.

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: A comparison of methods. *Statistics in Medicine*, *18*(20), 2693–2708.

Thornton, A., & Lee, P. (2000). Publication bias in meta-analysis: Its causes and consequences. *Journal of Clinical Epidemiology*, *53*(2), 207–216.

Tilling, K., Macdonald-Wallis, C., Lawlor, D. A., Hughes, R. A., & Howe, L. D. (2014).

Modelling childhood growth using fractional polynomials and linear splines. *Annals of Nutrition and Metabolism*, *65*(2-3), 129–138.

Tsui, A. S. M., Byers-Heinlein, K., & Fennell, C. T. (2019). Associative word learning in infancy: A meta-analysis of the switch task. *Developmental Psychology*, *55*(5), 934.

Tsuji, S., & Cristia, A. (2014). Perceptual attunement in vowels: A meta-analysis. *Developmental Psychobiology*, *56*(2), 179–191.

Vandenbroucke, J. P., Elm, E. von, Altman, D. G., Gøtzsche, P. C., Mulrow, C. D., Pocock, S. J., … Initiative, S. (2007). Strengthening the reporting of observational studies in epidemiology (STROBE): Explanation and elaboration. *Annals of Internal Medicine*, *147*(8), W–163.

Vevea, J. L., & Hedges, L. V. (1995). A general linear model for estimating effect size in the presence of publication bias. *Psychometrika*, *60*, 419–435.

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Visser, I., Bergmann, C., Byers-Heinlein, K., Dal Ben, R., Duch, W., Forbes, S., et al.others. (2022). Improving the generalizability of infant psychological research: The ManyBabies model. *Behavioral and Brain Sciences*, *45*.

Zettersten, M., Cox, C., Bergmann, C., Tsui, A. S. M., Soderstrom, M., Mayor, J., et al.others. (2024). Evidence for infant-directed speech preference is consistent across large-scale, multi-site replication and meta-analysis. *Open Mind*, *8*, 439–461.

Table 1

This table summarizes the number of effect sizes (ES) and the number of participants included in each dataset. The ES estimates represent the aggregated effect sizes and their 95% confidence intervals. The I^2 measures the heterogneity of each dataset. The paper source column indicates the published record associated with each dataset.

Dataset	N ES	N Subject	MA ES	I^2	Source paper / Data Curator
Abstract rule learning	95	1123	0.22 [0.07, 0.37]	0.80	Rabagliati et al., (2018)
Audio-visual congruence	92	4132	0.33 [0.19, 0.47]	0.89	Cox et al., (2022)
Categorization bias	80	594	0.16 [-0.66, 0.99]	0.96	Molly Lewis
Cross-situational word learning	48	2241	0.67 [0.5, 0.84]	0.90	Rodrigo Dal Ben
Familiar word recognition	34	586	0.54 [0.38, 0.69]	0.55	Carbajal et al., (2021)
Gaze following (combined)	81	1407	0.81 [0.61, 1.01]	0.90	Frank et al., (2016)
Infant directed speech preference	100	1267	0.37 [0.25, 0.49]	0.71	Zettersten et al., (2023)
Label advantage in concept learning	100	1644	0.36 [0.23, 0.48]	0.73	Molly Lewis
Language discrimination	104	1479	-0.26 [-0.4, -0.11]	0.77	Gasparini et al., (2021)
Language preference	49	641	0.11 [-0.06, 0.28]	0.93	Gasparini et al., (2021)
Mispronunciation sensitivity	249	2122	0.45 [0.24, 0.66]	0.94	Von Holzen & Bergmann (2021)
Mutual exclusivity	131	2222	1.27 [0.99, 1.56]	0.95	Lewis et al. (2020)
Natural speech preference	55	786	0.44 [0.23, 0.65]	0.83	Issard et al., (2023)
Neonatal Imitation	336	2455	0.68 [0.4, 0.97]	0.94	Davis et al. (2021)
Online word recognition	14	330	1.37 [0.78, 1.96]	0.95	Frank et al., (2016)
Prosocial agents	61	1244	0.4 [0.29, 0.52]	0.20	Margoni & Surian (2018)
Simple arithmetic competences	14	369	0.25 [0.04, 0.46]	0.54	Christodoulou et al., (2017)
Sound symbolism	44	425	0.16 [-0.01, 0.33]	0.69	Fort et al. (2018)
Statistical sound category learning	20	591	0.29 [0.01, 0.57]	0.58	Cristia (2018)
Statistical word segmentation	103	804	-0.08 [-0.18, 0.02]	0.83	Black & Bergmann (2017)
Switch task	143	2764	-0.16 [-0.25, -0.06]	0.78	Tsui et al., (2019)
Syntactic bootstrapping	60	832	0.24 [0.03, 0.44]	0.72	Cao & Lewis (2022)
Vowel discrimination (native)	143	2418	0.59 [0.43, 0.75]	0.78	Tsuji & Cristia (2014)
Vowel discrimination (non-native)	49	600	0.65 [0.2, 1.1]	0.92	Tsuji & Cristia (2014)
Word segmentation (combined)	315	2910	0.2 [0.14, 0.26]	0.78	Bergmann & Cristia (2016)

Table 2

This table summarizes the values of Δ of corrected Akaike Information Criterion (AICc) for the age model with different functional forms: Constant, Linear, Logarithmic, and Quadratic. The values were calculated from subtracting the minimum AICc from the AICc of each model. They were rounded to two decimals. Zeros represent the models with the best fit. The bold values indicate the best fitting model. Asterisks indicate that there is a significantly better fit compared to other functional forms for that dataset.

Dataset	Const	Linear	Log	Quadratic
Cross-situational word learning	**0.00**	2.44	2.29	2.55
Language discrimination	**0.00**	1.32	0.91	1.59
Prosocial agents	**0.00**	2.08	1.87	2.15
Simple arithmetic competences	**0.00***	6.65*	6.74*	6.55*
Statistical word segmentation	**0.00**	1.34	1.51	1.12
Switch task	**0.00**	1.12	1.15	1.06
Syntactic bootstrapping	**0.00**	0.71	0.56	0.88
Vowel discrimination (native)	**0.00**	1.34	0.99	1.63
Vowel discrimination (non-native)	**0.00**	1.56	1.67	1.46
Word segmentation (combined)	**0.00**	1.28	1.05	1.61
Infant directed speech preference	**0.00**	1.57	1.47	1.53
Mispronunciation sensitivity	1.89	**0.00**	0.05	0.19
Online word recognition	2.22	**0.00**	0.23	0.15
Sound symbolism	3.91	**0.00**	0.61	0.09
Audio-visual congruence	5.90*	6.70*	**0.00***	7.44*
Label advantage in concept learning	2.37	0.95	**0.00**	1.63
Mutual exclusivity	9.80*	0.58	**0.00***	1.38
Neonatal Imitation	2.25	0.36	**0.00**	1.06
Abstract rule learning	0.44	0.32	0.86	**0.00**
Categorization bias	8.46*	0.62	1.36	**0.00***
Familiar word recognition	1.68	0.28	1.15	**0.00**
Gaze following (combined)	43.73*	2.07	10.41*	**0.00***
Language preference	2.50	2.36	4.12	**0.00**
Natural speech preference	0.86	0.43	1.04	**0.00**
Statistical sound category learning	3.44	1.04	3.01	**0.00**

Table 3

This table presents whether the original dataset shows any evidence for linear growth, and to what extent there is evidence supporting the four hypotheses (Checkmarks for yes, crosses for no). Absence of any symbol suggests that there is not enough data to test the hypothesis.

Dataset	Linear Growth	H1		H2	H3	H4
		Weight Function	Egger's Test			
Abstract rule learning	X	X	X	X		
Audio-visual congruence	X	X	✓	X		X
Categorization bias	X	X	✓			X
Cross-situational word learning	✓		X	X		✓
Familiar word recognition	✓	X		X	X	
Gaze following (combined)	✓	X	X			✓
Label advantage in concept learning	X	X	X			X
Language discrimination	X		X	X		
Language preference	X	X	X	X		
Mispronunciation sensitivity	✓	X	X			✓
Mutual exclusivity	✓	✓	X	X	✓	✓
Natural speech preference	X	X	X	X		
Neonatal Imitation	✓	X	X			
Online word recognition	✓			✓		✓
Prosocial agents	X		X	X	X	X
Simple arithmetic competences	X					
Sound symbolism	✓		X	X	X	X
Statistical sound category learning	✓			✓	✓	
Statistical word segmentation	X	X	X	X	X	
Switch task	X	X	X	X	X	X
Syntactic bootstrapping	X	X	✓	X	X	X
Vowel discrimination (native)	X	X	X	X		X
Vowel discrimination (non-native)	X	✓	X	X		
Word segmentation (combined)	X	X	X	X		X

Table 3 continued

Dataset	Linear Growth	H1		H2	H3	H4
		Weight Function	Egger's Test			
Infant directed speech preference	X	X	X	X		X

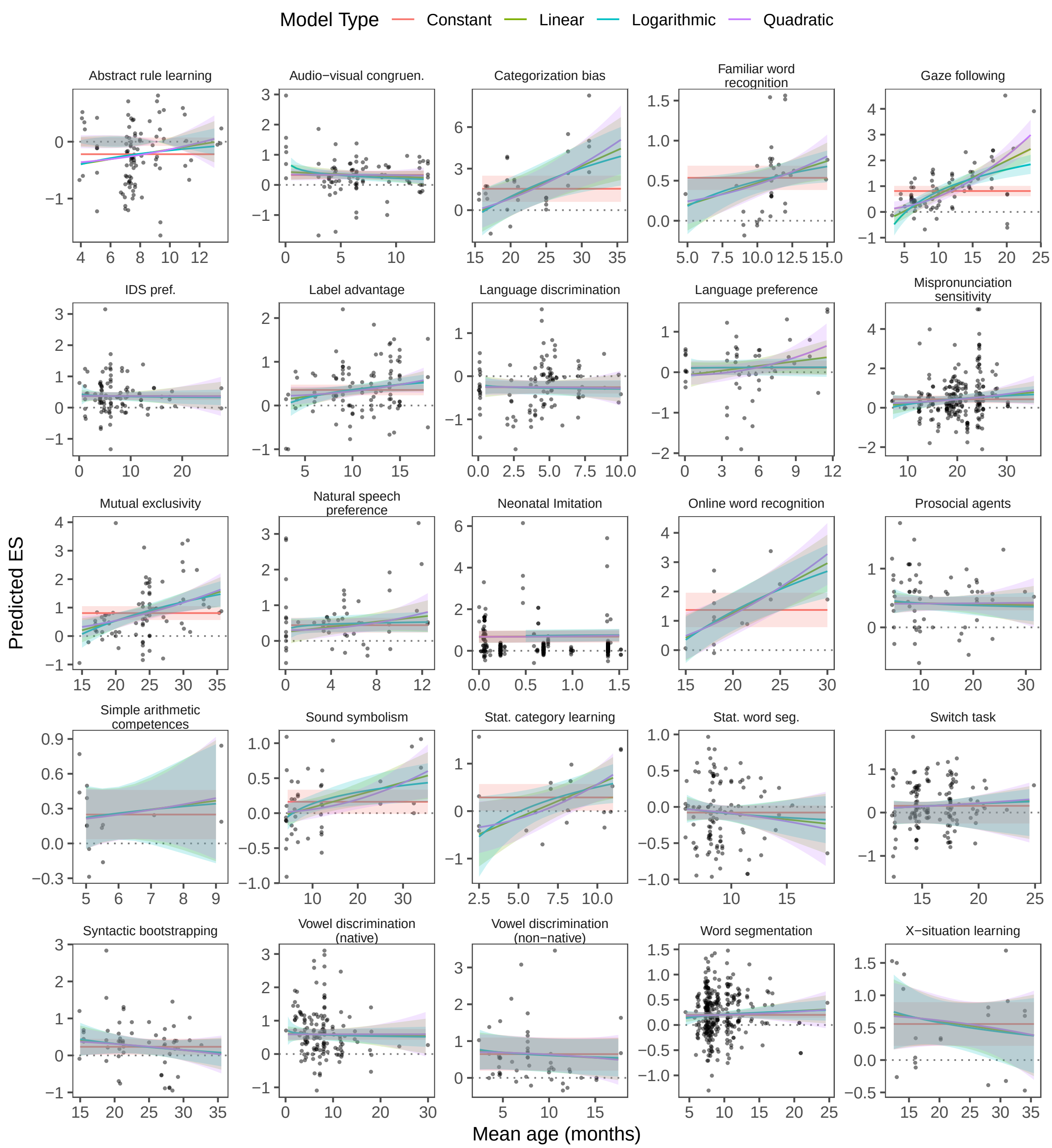

Figure 1. Each panel shows the dataset and the predicted values of the four functional forms. For each panel, X-axis represent the age in month, and Y-axis represents the effect size. The shaded area is the 95% confidence interval of the prediction.

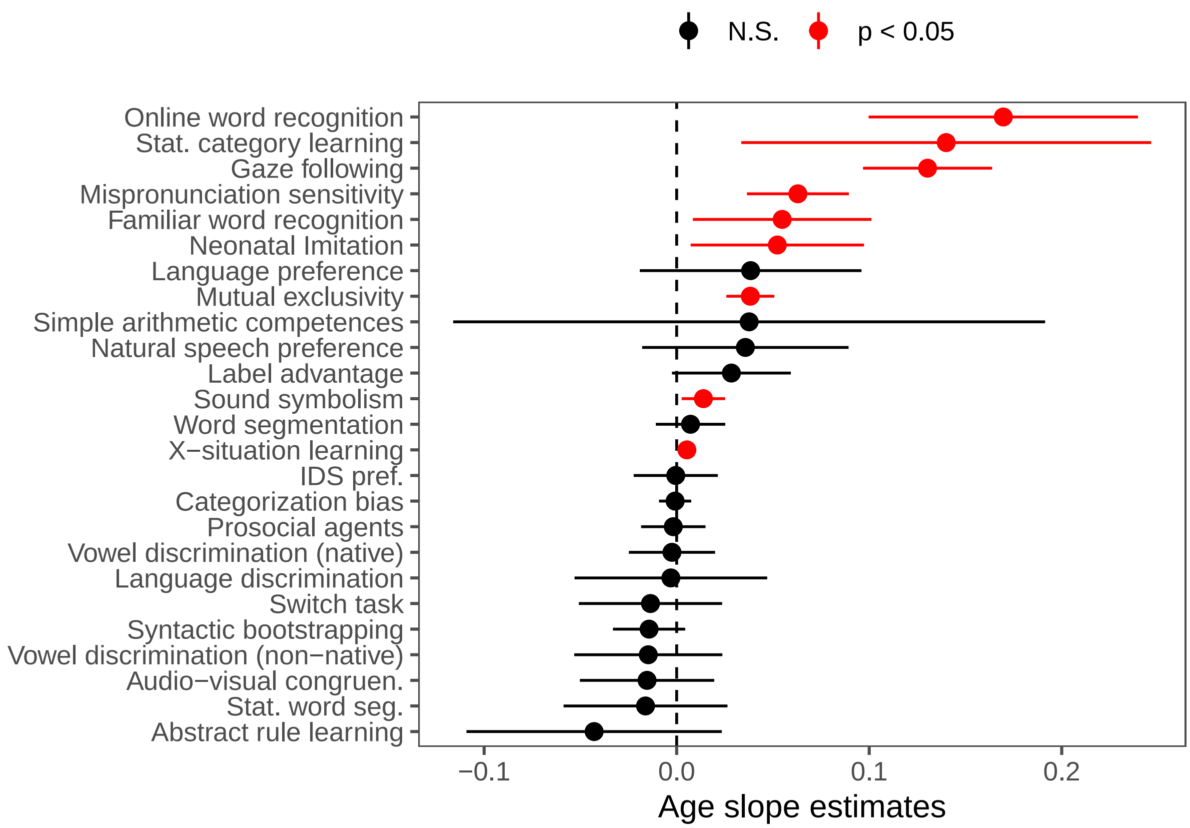

Figure 2. Each dot represents the estimate of the age predictor in the linear model. Red dotsindicate the particular estimate is statistically significant, and black indicate the estimate is not significant. Error bars show 95% confidence intervals.

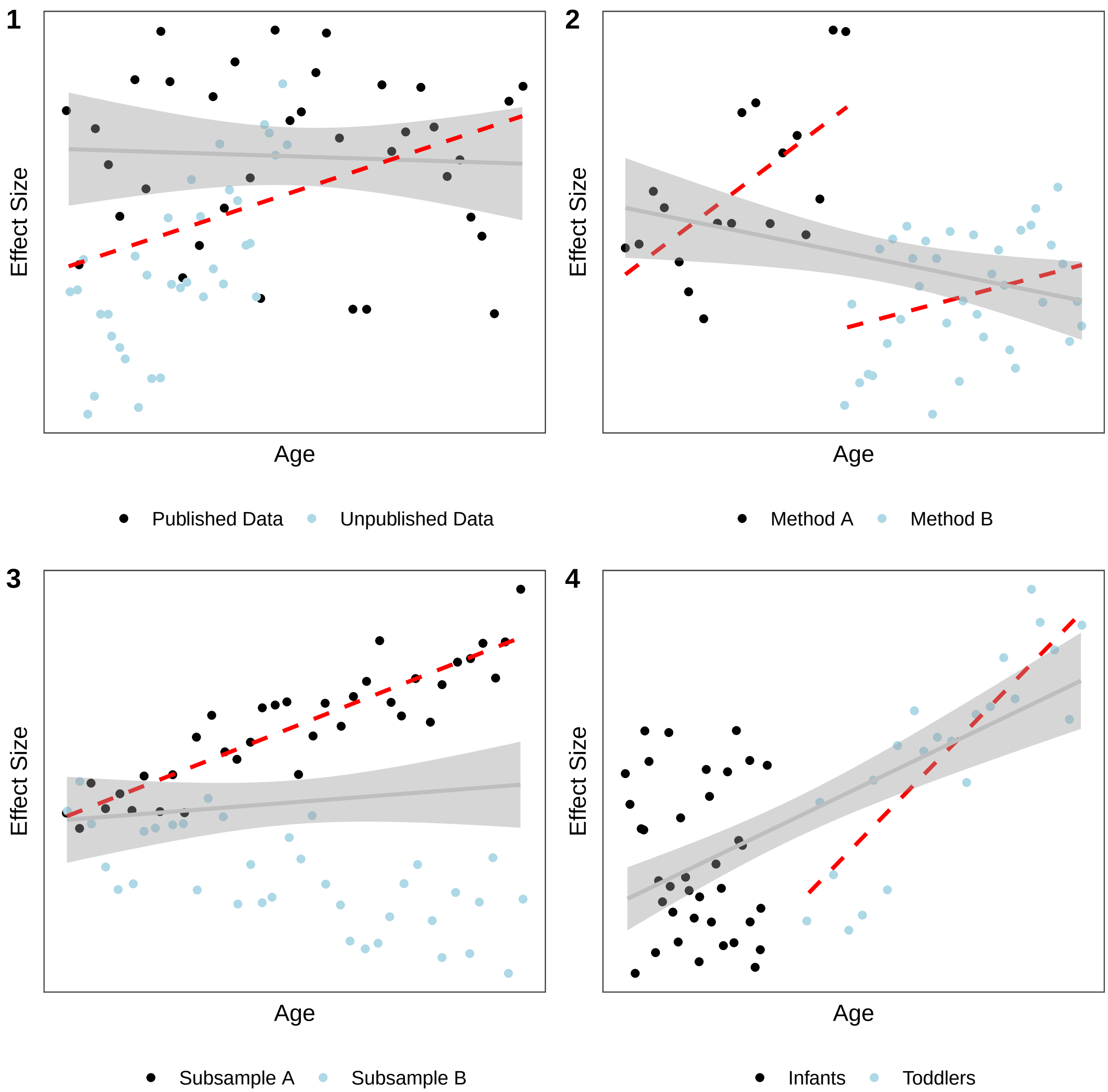

Figure 3. Schematic illustration of the four hypotheses considered. The gray shaded line represents the observed age effect if the hypothesis holds. The red dotted lines represents the true underlying age effect under the particular hypothesis.