Journal of Child Language (2025), 1–36
doi:10.1017/S0305000924000692

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

How strong is the relationship between caregiver speech and language development? A meta-analysis

Joseph R. Coffey and Jesse Snedeker

Department of Psychology, Harvard University, Cambridge, MA, USA
Corresponding author: Joseph R. Coffey; Email: jrcoffey@g.harvard.edu

(Received 07 January 2024; revised 15 November 2024; accepted 01 December 2024)

Abstract

A growing body of research has found that talking to young children is positively associated with language outcomes. However, there is tremendous heterogeneity in the design of these studies, which could potentially affect the strength and reliability of this association. The present meta-analysis, comprising 4760 participants across 71 studies, goes beyond prior research by including: 1) more recent studies, 2) non-English-speaking populations, 3) more fine-grained categorization of measures of input, 4) additional moderators, and 5) a multi-level model design allowing us to consider multiple effect sizes per study. We find a moderate association between input and outcomes (R^2=0.04-0.07) across four input measures, with some evidence of publication bias. We find no differences in effect size across any of the input measures. Child age and study duration moderated some effects of input. Our findings suggest that language input-outcome associations remain robust but modest across a multitude of contexts and measures.

Keywords: meta-analysis; language development; publication bias; culture; SES

1. Introduction

Children acquire the languages used by those around them. Toddlers in English-speaking families say *dog*, while those in French-speaking families say *chien*. Thus, all theories of language development are grounded in the assumption that the language children experience (i.e., the input) plays a critical role in language development. Over the last forty years, however, a substantial body of research has been conducted to test a stronger claim: that differences in the amount or kind of input that children receive are associated with differences in the pace with which they acquire language (Zauche et al., 2016).

In these studies, researchers record interactions between a young child and their caregiver during daily activities (like meals) or a structured play session. This speech sample is then transcribed and coded for properties that might predict individual differences in children's language development, such as the amount of speech directed at the child (Huttenlocher et al., 1991), its lexical diversity (Hart & Risley, 1995), or grammatical complexity (Hoff-Ginsberg, 1986). Finally, parent input is compared to the child's language ability, which can be assessed in a wide variety of ways, such as by administering a vocabulary assessment, collecting the information via a parent report, or sampling speech produced by the child during observation. Most of these studies find that parent input predicts their children's language outcomes. This work has been cited as motivating researchers to develop interventions that seek to increase parents' verbal engagement with their children (e.g., Dupas et al., 2023; Suskind et al., 2016; Weber et al., 2017; Wong et al., 2020).

The present paper is a meta-analysis exploring the size of these input–output correlations, the range of conditions under which they are observed, and the degree to which the size of these effects depends on the choice of input measure, outcome measure, study design, or population studied. In the remainder of this introduction, we review the principal findings of the input literature, describe meta-analytical methods and what they can accomplish, present the questions motivating the present meta-analysis, and discuss the findings from previous meta-analyses on this topic and how our work goes beyond this study.

1.1. The growing interest in studies of caregiver input

While research into the relationship between parental speech and language development began in the 1970's (e.g., Newport & Gleitman, 1977), much of the current interest in this topic stems from Hart & Risley's 1995 book (H&R). H&R followed 42 families for nearly 2.5 years, collecting data on children's linguistic milestones and sampling naturalistic speech in the home during monthly hour-long visits. They found that the thirteen children in the highest income group heard on average over 2000 words per hour from their primary caregiver, whereas the six children in families receiving welfare heard around 600. Critically, differences in caregiver input predicted individual differences in the rate of children's vocabulary growth, such that children who heard more words had larger vocabularies than children who heard fewer. These effects persisted: parental vocabulary use at age 3 predicted performance on standardized language tests at age 9 (Walker et al., 1994). The authors concluded that child-directed speech plays a critical role in language development.

While more recent studies have found that socioeconomic differences in child-directed speech are neither as large nor as prevalent as H&R suggest (e.g., Dailey & Bergelson, 2022; Sperry et al., 2019), a growing body of research has supported H&R's second conclusion that differences in the properties of caregiver input predict differences in language growth. These findings have been replicated in a variety of socioeconomic contexts (e.g., Hoff, 2003; Pan et al., 2004; Rowe, 2008; Huttenlocher et al., 2010; Hirsh-Pasek et al., 2015; Romeo et al., 2018). While most input studies have been conducted with English-speaking families in the U.S., similar patterns have also been observed in other contexts (e.g., Hurtado et al., 2008; Mastin & Vogt, 2016; Shneidman & Goldin-Meadow, 2012; Weber et al., 2017; Weisleder & Fernald, 2013; Zhang et al., 2023).

1.2. Motivating the meta-analytic approach

While there is a broad consensus that input and outcome are correlated, several important questions remain that can be addressed by meta-analysis. First, it is unclear how large these correlations are. H&R found that input measures collected between 34 and 36 months of age predicted over half of the variance in vocabulary at 36 months

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

(R^2 = 0.53), suggesting that input variation is the primary factor setting the pace for vocabulary growth. Studies conducted since then, however, have found a wide range of effect sizes ranging from R^2 = 0.61 (Leech & Rowe, 2014) to R^2 = 0.00 (Pancsofar & Vernon-Feagans, 2006). Knowing how large we should expect these effects to be, in general, allows us to more accurately determine how typical or unusual a given finding is, opening the way for further discovery. This information also allows researchers to set their priors more accurately for power analyses. Finally, it allows for direct comparisons of parental input to other predictors of language development (e.g., Is parent input *more* predictive of language outcomes than input from teachers or genetic differences that impact learning?).

Second, there is mixed consensus on which input measures are *most* predictive of language outcomes. Child-directed speech is a rich stimulus that can be characterized in a variety of ways. Researchers may be interested in whether simply hearing more speech facilitates development or whether the benefits come from hearing speech of a specific *kind*. Thus, speech coding is often broken down into two categories: the *quantity* and *quality* of the input. *Quantity* measures, like the number of words, capture how much speech children hear during interactions with their caregivers. *Quality* measures, like lexical richness, capture the degree to which caregiver speech contains features that are thought to facilitate language learning. Some researchers have found that measures of quality are more associated with child outcomes than measures of quantity, particularly later in development (e.g., Hsu et al., 2017; Pan et al., 2004; Rowe, 2012). Findings of this kind are central to understanding the mechanism by which input shapes language development and the kind of input that matters most (Golinkoff et al., 2019). Thus, it is important to know if such patterns are consistently observed across studies and how large the difference in effect size is. In our study, we focus on the four measures that are most often reported in input studies. Two are measures of *quantity* (number of utterances and word tokens), and two are measures of *quality* (number of word types and the mean length of utterances). Word tokens are counted as the number of words in the sample, while word types are counted as the number of different words in the sample. Mean length of utterance, or MLU, is defined as the average length of an utterance in words or morphemes.

Third, because input studies differ greatly from one another, differences in study design, participant characteristics, or language measures could moderate the effects of input. Speech from parents can be sampled: in the participants' homes or in the researcher's laboratory; during episodes of play or reading; for a few minutes or for many hours. Children's language ability can be assessed via direct testing, by observing children's speech production with their caregivers, or by surveying their primary caregiver about their language use. Studies also vary in the characteristics of the participants, such as the age of the child or the gender of the parent. Studies like H&R that find correlations between SES and input raise the possibility that the strength of the association between input and outcome might depend on the SES composition of the sample, with more economically diverse samples having larger correlations.

Finally, there is a possibility that research on input–outcome relationships is skewed by publication bias. The association between children's language and caregiver speech is supported by a large literature and has become a fixture of interventions and policy initiatives (Dupas et al., 2023; Suskind et al., 2016; Weber et al., 2017; Wong et al., 2020). Thus, many researchers may expect to find such a relationship in their data. This kind of consensus can create a "file-drawer effect," where negative or null results are not published (Rosenthal, 1979), which would inflate the apparent effect size. Fortunately, meta-analysis offers tools to estimate the possible effects of publication bias. By comparing data that has been published to unpublished data obtained through contact with

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

authors, one can determine whether positive results are more likely to be reported than null results. One can also create funnel plots, which visualize the distribution of effect sizes relative to their standard errors, to determine whether there are more positive results than would be expected in studies with greater variance (suggesting selective reporting). Methods like Egger's test can be used to determine whether effect sizes are asymmetrically distributed against standard errors (Egger et al., 1997). These plots can also include contour lines plotting the distribution of p-values, which can reveal other possible causes of asymmetry, such as variable study quality (Peters et al., 2008).

1.3. Prior meta-analyses

Two prior meta-analyses have explored input–outcome correlations. Wang et al. (2020) conducted a meta-analysis comparing language outcomes to caregiver input collected via the use of LENA recording devices, which produce automatic quantitative measures of speech spoken near the child extracted from hours of audio. They collected 17 studies of children from birth to 48 months, including two non-English studies (one in Mandarin and one in Finnish), and ran three analyses on each of the LENA input measures: adult word counts, child vocalization counts, and conversational turn counts. Collapsing across measures, they found a moderate relationship between LENA measures and language outcomes ($R^2 = 0.07$). Examining each measure individually, they found that adult word counts had the weakest relationship with outcomes ($R^2 = 0.04$), followed by conversational turn counts ($R^2 = 0.09$) and child vocalizations ($R^2 = 0.09$). In addition, longer elapsed time between input and outcome collection was associated with smaller correlations. No evidence of publication bias was found.

Most recently, Anderson et al. (2021, AGPJM) conducted a meta-analysis focusing on the relationship between language outcomes and input *quantity* and *quality*. Studies were included in the analysis of *quantity* if they included either word tokens or utterances as an input measure. Studies were included in the analysis of *quality* if they included either a measure of input *diversity* (e.g., word types/roots, type-token ratio) or *complexity* (mean length of utterance, rare word usage, lexical richness, or multi-clausal utterances). Only studies examining typically developing English-speaking children were included, and studies using LENA were omitted. The analyses included 33 quantity studies and 35 quality studies of 1- to 72-month olds. Using hierarchically ordered study selection criteria, they selected a single statistical effect size from each study. Studies of quantity had significant effects overall ($R^2 = 0.04$). There were two reliable moderators: effect sizes were larger in longitudinal studies and in studies where children were assessed in naturalistic contexts. The funnel plot was asymmetric for the quantity studies, suggesting publication bias, but the authors estimated that the true effect size was only slightly smaller than the reported effect size. Studies of quality had larger effect sizes ($R^2 = 0.11$) with no evidence of publication bias. For studies of quality, there were larger effect sizes for longer studies, studies where input was collected in naturalistic contexts, and studies of older children. They conducted separate analyses on their measures of *diversity* (primarily types, $R^2 = 0.07$) and *complexity* (primarily MLU, $R^2 = 0.11$), which were not significantly different from one another.

1.4. Current study

This study builds upon the AGPJM meta-analysis but goes beyond it in six critical ways.

First, because this research area is highly active, we were able to include 23 studies that were not available when AGPJM conducted their analysis.

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

Second, we adopted a multilevel mixed-effects design that allows us to include multiple effect sizes per study (e.g., correlations collected with different language assessments or different speakers), increasing the total number of effect sizes analysed from k = 71 to k = 323. By using *robust variance estimation* to control for statistical interdependence between effect sizes in the same study, we are able to include all effect sizes reported in our sample, increasing the precision of our estimates (Pustejovsky & Tipton, 2022). This approach was also used by Wang et al. (2020) and departed from our original preregistered analysis, which used hierarchical study selection to derive a single effect size per study. The findings of the preregistered analysis, which differed little from the current analysis, can be found in the Supplementary Materials.

Third, we adopted a finer-grained coding system for input, examining the four standard measures of input separately: utterances, tokens, types, and MLU. This approach differentiates between measures of input within *quality* and *quantity* and allows us to perform more precise analyses of pooled effect size and moderators. Using multilevel modelling, we could also compare the pooled effect sizes of these different measures directly by including them in the meta-regression model as moderating variables. We use a similar technique to test for publication bias. If there were publication bias, we would expect studies with large standard errors to have disproportionately large and positive effect sizes. We can test for this by constructing a regression model with standard error as a moderating variable (Rodgers & Pustejovsky, 2021).

Fourth, in addition to exploring previously studied moderators using our novel statistical approach, we included several moderators that have gone unexamined. For example, we were interested in whether effect sizes are larger when input and outcome measures are identical, which often occurs when children's language is assessed on the basis of the input speech sample. If the context of the conversation shapes the behaviour of both participants in similar ways (e.g., a discussion of different animals might produce greater lexical diversity from parent and child than pretend play with cars), we might expect these studies to produce larger correlations.

Fifth, we wished to investigate whether input–outcome associations showed evidence of boundedness. Most studies of input claim that associations between input and outcomes are causal in nature: children who hear speech more often or hear more complex speech learn their language more quickly. To our knowledge, no studies have examined whether the effects of input are bounded. In other words, is there a point at which children begin seeing diminishing returns on learning from additional input? Intuitively, we might expect there to be a limit to how much new information young children can learn in a day, after which additional child-directed speech does not produce more learning. If this were true, we would expect input–outcome correlations to be smaller in samples where children, on average, hear more speech or more complex speech.

Finally, we have taken a simple step to explore the effects of language and culture on input–outcome associations by including non-English studies in our analysis. While we do not have access to data from studies of rural, non-Western populations, we wanted to determine how large effect sizes were in non-English and in non-U.S. samples, and whether they differed substantially from English/U.S. samples. This analysis was not included in the meta-analyses reviewed earlier. This analysis is an initial, small step towards understanding whether there are systematic differences in these effects across languages and countries.

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

2. Methods

We followed PRISMA guidelines for conducting and reporting results for meta-analyses (Page et al., 2021) (Figure 1). A full breakdown of our study selection procedure can be found in the Supplementary Materials.

2.1. Search procedure

2.1.1. Forward search

We constructed a Boolean search query based on our inclusion criteria, searching the abstract, title, and keywords for references to (1) caregiver participants, (2) child participants, (3) input measures, and (4) output measures. We refined our query by testing candidate queries to ensure that the 28 studies in our literature review (see below) were found. We surveyed the following databases based on their relevance to research in child language development: ERIC, PsycInfo, Academic Search Premier, PubMed, Web of Science, and Proquest (for dissertations). The most recent search was conducted 14 January 2021 and produced 6763 abstracts for further screening.

2.1.2. Other sources

Expert knowledge: Prior to our search, we included 28 publications for eligibility drawn from a literature review conducted for a previous study of child-directed speech (Coffey et al., 2022).

Contacting authors: We reached out to the research community through the ICIS and CHILDES email listservs for missed studies and unpublished data. We considered an additional eight publications collected this way.

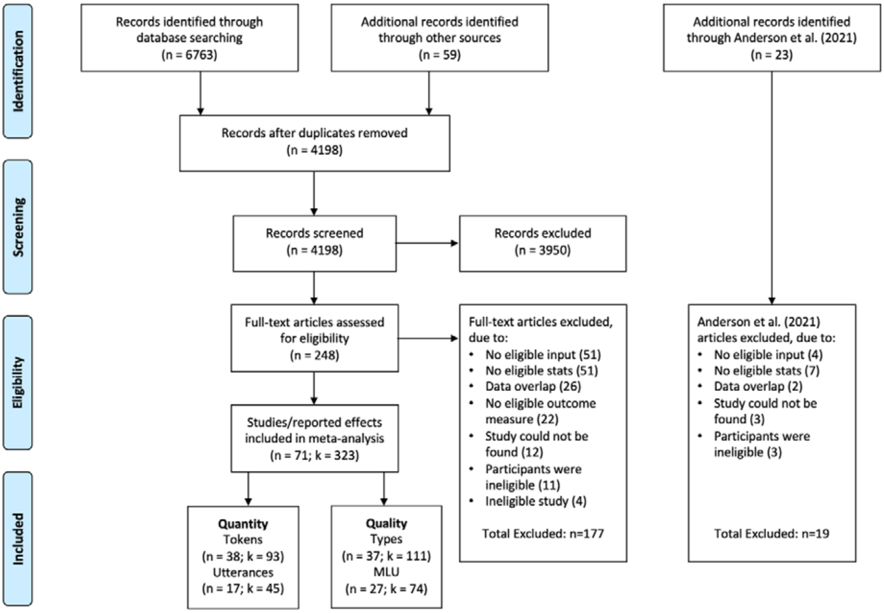
Identification
Screening
Eligibility
Included
Records identified through database searching (n = 6763)
Additional records identified through other sources (n = 59)
Additional records identified through Anderson et al. (2021) (n = 23)
Records after duplicates removed (n = 4198)
Records screened (n = 4198)
Records excluded (n = 3950)
Full-text articles assessed for eligibility (n = 248)
Full-text articles excluded, due to:
• No eligible input (51)
• No eligible stats (51)
• Data overlap (26)
• No eligible outcome measure (22)
• Study could not be found (12)
• Participants were ineligible (11)
• Ineligible study (4)
Total Excluded: n=177
Anderson et al. (2021) articles excluded, due to:
• No eligible input (4)
• No eligible stats (7)
• Data overlap (2)
• Study could not be found (3)
• Participants were ineligible (3)
Total Excluded: n=19
Studies/reported effects included in meta-analysis (n = 71; k = 323)
Quantity
Tokens (n = 38; k = 93)
Utterances (n = 17; k = 45)
Quality
Types (n = 37; k = 111)
MLU (n = 27; k = 74)

Figure 1. PRISMA flowchart detailing study collection.

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

Prior meta-analysis: Finally, we compared the results of our literature review, forward search, and author contacts to the list of publications included in the AGPJM meta-analysis. We considered an additional 23 publications collected this way. Nineteen were excluded due to data overlap or differences in our respective inclusion criteria (Figure 1).

2.2. Inclusion criteria

We screened study abstracts for four criteria. First, studies must be English-language journal articles, book chapters, dissertations, or conference proceedings. Reviews and meta-analyses were excluded. Second, studies must examine typically developing, monolingual children between the ages of 1–8. Atypically developing children, such as children with autism or preterm infants, were excluded, as well as multilingual children. Third, studies must include one of our four input measures (utterances, tokens, types, or MLU) from speech directed to children by caregivers in naturalistic or semi-naturalistic settings. Studies that only collected other measures of input (e.g., questions, decontextualized speech) or measures of interaction (e.g., warmth, responsiveness) were excluded. We also excluded input that was scripted (e.g., only words read from a book). Fourth, studies must include either: a measure of children's vocabulary, a broad measure of language development, or an observation of children's language use. We excluded studies that measured other specific forms of language proficiency (syntactic knowledge, pragmatics, novel word-learning) primarily because of the small number of studies using any one of these measures. Our screening of abstracts left 248 potentially eligible publications. We located full-text versions of these studies for further review and coding. Studies with data that reported significant overlap with other studies were reviewed further. When available, the earliest reporting of the original data set was preferred. We assumed the initial reporting of data would be the primary analysis, whereas subsequent reports would be affected by what had been found previously. Full-text review was conducted by the lead author.

This procedure resulted in 75 studies that were coded for effect size and moderating variables. Four studies that did not provide Pearson's *r* were omitted from analysis. In addition, five reported input–outcome correlations were calculated from composited measures of input (e.g., taking the average of multiple standardized input values) and could not be included in any of the individual input analyses. In total, we examined 71 studies, reporting 323 correlations, across 4760 unique participants.

2.3. Study coding

Coding was conducted by the lead author. An additional annotator independently coded 21 studies to check for accuracy. Studies were coded four different kinds of variables: input measures, outcome measures, subject characteristics, and study characteristics.

2.3.1. Input measures

Word tokens: Word tokens are a raw count of the total number of words produced. Tokens produced by parents have been found to predict language development in children (e.g., Hoff, 2003; Rowe, 2012).

Total utterances: Utterances are defined as a continuous segment of speech. An utterance can be a single sentence, a word, a phrase, or a portion of a sentence and are commonly delimited by pauses in speech. The number of parent utterances have been found to predict individual differences in language outcomes (e.g., Pancsofar & Vernon-Feagans, 2006).

Word types: Word types are a measure of how many different words are produced in a sample. They index the lexical diversity of speech or the number of different words that children have the opportunity to learn. Parent word types are frequently found to predict language outcomes (e.g., Hart & Risley, 1995; Rowe, 2012). In addition to word types, we also included closely associated measures, like number of different word roots or morphemes.

Mean-length of utterance (MLU): MLU is the average number of linguistic units an utterance contains. This measure has been used to index the grammatical complexity of an utterance. MLU and language outcomes are often positively correlated (e.g., Hoff-Ginsberg, 1986). MLU is commonly defined as the average number of morphemes in an utterance but can also be defined as the average number of words in an utterance. These measures are highly correlated (Parker & Brorson, 2005), and thus, we included both in our analysis.

2.3.2. Outcome measures

We coded studies for how they measured language outcomes. We distinguished between three kinds of assessment types: observation, direct assessment, and parent report. Within these assessments, we also distinguished between two kinds of assessment measures: expressive and receptive language. We also distinguished whether an assessment was a measure of vocabulary specifically. Finally, for studies using word types as the input measure, we coded whether the outcome measure captured the same construct in the child (e.g., parent word types and child word types produced during observation). Studies that did were coded as matched.

2.3.3. Subject characteristics

Gender was coded as the percentage of children who were female. Age was coded at the time input measures were collected and the time outcome measures were collected. We coded information about the speaker(s) providing input (e.g., mother, father, primary caregiver, etc.) and the native language and country of origin of the household. Next, we coded for household SES by categorizing studies into three groups used by AGPJM. Studies focusing on samples with lower levels of income/education (relative to national norms) were coded as low SES. Studies with samples from across different income/education levels were coded as diverse SES. All other studies were coded as middle/high SES by default. Finally, for each input measure collected in a study, we coded mean input, or the average recorded value across observations. To ensure input means were comparable from study to study, we normalized word tokens and utterances for observation duration, producing measures of word tokens per minute and utterances per minute, respectively. This was unnecessary for MLU, which is already normalized for total caregiver utterances. We did not code the average word types produced because number of word types declines as a function of time (i.e., the longer a session goes on, the less likely new word types are to be encountered), and thus, word types per minute is confounded with observation duration.

2.3.4. Study characteristics

Studies were coded for their total language sample duration in minutes. Study location was coded as home, lab, or other. The activity taking place during the study was coded as

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

either naturalistic (participants were told to go about their day), semi-naturalistic play (participants were asked to play as they would normally), structured play (participants were asked to play with a particular set of toys), or other. Studies were also coded for temporal design, either cross-lagged or concurrent. Cross-lagged studies were those for which parent input was collected at a different time (Time 1) than children's outcomes (Time 2). Finally, we coded publication type. First, we coded a baseline set which included all studies that were published in peer-reviewed journals and where our correlation coefficient was taken from that paper. Next, data from books, dissertations, or other non-peer-reviewed sources were coded as non-peer-reviewed. Then, all studies in which the correlation coefficient was not included in the paper but could be calculated from the paper or were retrieved after contacting authors were coded as non-reported. We conducted two moderator analyses: one comparing baseline studies to non-peer-reviewed studies and another comparing baseline studies to studies coded as non-peer-reviewed or non-reported.

2.3.5. Effect sizes

Studies were coded for Pearson's r correlations between input and outcome measures. When these measures could not be found in the study, we reached out to the authors for either the correlation coefficient or the raw data from which we could calculate the correlation ourselves. We did not include partial correlations or regression coefficients with covariates to maintain the comparability of effects across studies. All correlation coefficients were normalized via conversion to z-scores and division by the squared inverse of their standard errors, or $\frac{1}{n-3}$, where n is the sample size for each study (Hedges & Olkin, 1985).

2.4. Analysis

All analyses were conducted in R (v 4.4.1) (R Core Team, 2024), using the *metafor* (v 4.6.0) and clubsandwich (v 0.5.11) packages (Pustejovsky, 2024; Viechtbauer, 2010). To determine the size of input–outcome correlations, random effects meta-regression models were fitted for each of the four input measures, where the intercept is the pooled effect size estimate. These models controlled for interdependence between shared effect sizes using robust variance estimation (*rho* = 0.8). Q-statistics from the resulting models were used to assess whether there was sufficient heterogeneity across studies to motivate an analysis of potential moderating variables. Moderators were then added to each of the base models as predictors. When considering continuous moderators, we removed outliers by excluding studies that reported values more than three standard deviations from the mean in our sample to avoid data skew. These studies were included in all other analyses. Statistical significance was determined in a two-step process: first, individual coefficients are tested for significance in the regression; second, likelihood ratio testing is used to determine whether the addition of the variable improved fit from the base model. When multiple moderators were found to be significant alone, they were included together in a single model, which was then checked for significance and improvement of fit. Base cases for categorical moderators were dummy coded zero and indicated in each table (e.g., for household SES, the base case is middle-upper or MU).

We then checked for publication bias using two methods. First, we publication status as a binary predictor variable in our moderator analysis. Second, we constructed funnel

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

plots for each of our input variables. These figures plot the correlation reported for each study against its standard error. Studies with smaller standard errors would be expected to find correlations closer to the pooled effect size than studies with larger standard errors (resulting in a downward funnel). If there were publication bias, it would lead to the disappearance of studies with large standard errors but small (non-significant) effect sizes, resulting in an effect size that gets larger as the standard error increases. To test for this, we approximated Egger's regression test for asymmetry in *metafor* by constructing a regression model using the standard error of each effect as a moderator (Egger et al., 1997; Rodgers & Pustejovsky, 2021). We then applied our criteria for moderator significance (regression and likelihood ratio test) to determine whether there was a significant effect of standard error on effect size.

Finally, we determined whether certain forms of input were more strongly associated with children's outcomes by fitting a single multilevel model with all data from all input types, including input measure as a moderator and applying our significance criteria.

3. Results

Our preregistration, data, code, and Supplementary Materials can be accessed through OSF (https://osf.io/aydcf/). A version of this analysis using hierarchical study selection, replicating most of the findings below, can be found in the Supplementary Materials.

A full breakdown of the study characteristics for each input analysis is given in Table 1. To determine whether the studies using the different input variables differed along other dimensions, we conducted a series of mixed-effect linear regressions using *lme4* (v 1.1-35.4) (Bates et al., 2015), where the moderating variable is used as the response variable,

Table 1. Summary of average study characteristics across all input analyses (median and range given for continuous variables)

	Tokens	Utts	Types	MLU
Studies (N)	38	17	37	27
Reported Effects (k)	93	45	111	74
Participants (n)	1986	956	2420	2340
Input (per min)	46 (6.11–108)	14.76 (4.57–50.76)	-	3.81 (2.48–6.28)
Duration (min)	26.5 (2.3–1920)	12.5 (4.73–43)	15 (2.3–1740)	15 (2.3–60)
Gender (%)				
Female	0.49	0.48	0.5	0.48
Male	0.51	0.52	0.5	0.52
Age (months)				
T1 - Input	20.13 (7–69.6)	16 (5–60)	23.91 (1–67)	20.74 (6–58.72)
T2 - Outcome	24.3 (12.37–91)	24 (12–62)	25 (9–91)	28 (12–58.72)
Region				
US	30	11	30	21
Non-US	8	6	7	6

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

Table 1. *(Continued)*

	Tokens	Utts	Types	MLU
Language				
English	30	12	31	22
Non-English	8	5	6	5
Household SES				
Middle-Upper	20	13	24	16
Diverse	10	3	6	8
Low	8	1	7	3
Language assessment				
Direct assessment	17	6	14	11
Parent report	16	10	13	11
Observed	14	5	21	16
Composite	0	1	1	1
Language measure				
Expressive	30	14	30	23
Receptive	14	9	13	8
Both	3	1	3	4
Input/Outcome relation				
Cross-lagged	24	10	21	16
Concurrent	22	10	24	15
Parent–Child activity				
Naturalistic	18	3	10	4
Nat. Free play	6	5	4	6
Strc. Free play	11	7	17	12
Other	6	2	9	7
Recording method				
Video	21	14	29	20
Audio	4	2	5	6
LENA	13	1	3	1

input measure is included as a categorical predictor (word tokens coded as zero), and study ID is used as a random intercept. We used mixed-effect modelling for the comparisons to account for multiple values introduced by the same study. We found that studies of word types were shorter on average than studies of word tokens ($\beta = -15.86$, $SE = 7.16$, $p = 0.03$). Studies of word tokens were longer than other studies, but these comparisons did not reach significance.

We provide forest plots for each analysis: for studies with multiple reported correlations, a single composited correlation was calculated using *aggregate* using *metafor*. Reported pooled effect sizes have been converted from the *z*-score to *r* for interpretability. Moderator effects are given in tables, with significant effects bolded (i.e., significant coefficient and improved model fit by χ^2 test). Finally, funnel plots are presented illustrating each reported correlation plotted against its standard error, its statistical significance, the overall pooled effect size (with 95% confidence intervals), and the degree of distributional asymmetry given by an Egger's test.

3.1. Word tokens

We examined 93 correlations across 38 studies that measured word tokens (n = 1986 unique participants). The number of effect size estimates per study ranged from 1 to 6 (median: 2). We found a medium-sized effect across studies (r = 0.23, Figure 2). Q-statistics revealed significant evidence for between-study heterogeneity (Q(92) =362.68, p < 0.001), which motivated an analysis of possible moderators. While some moderators were significant when included in the model (Table 2), none of them resulted in improved model fit.

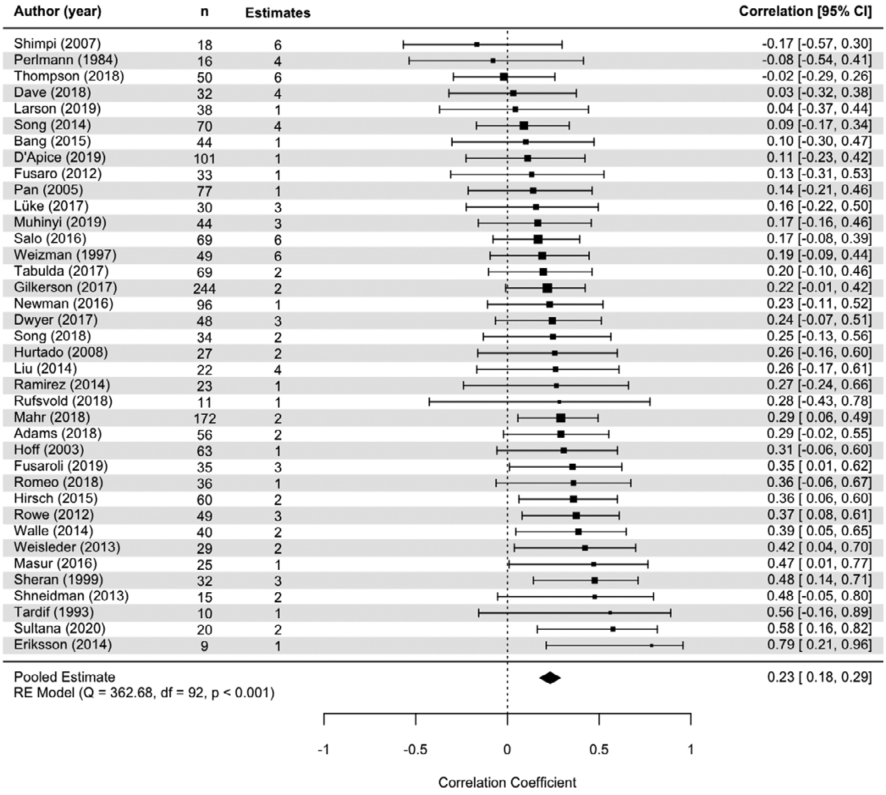

Figure 2. Forest plot of token study correlations.

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

Table 2. Summary of moderator analysis for studies of word tokens

Moderator	n	k	intcpt	β (SE)	p_β	AICc	χ^2	p_χ
Base model	38	93	0.24			−24.45		
Subject characteristics								
Child Gender (% Female)	36	89	0.23	0.01 (0.02)	0.50	−22.51	0.25	0.62
Household SES (MU)	38	93	0.22			−23.29	1.91	0.38
versus Diverse				0.08 (0.06)	0.19			
versus Low				−0.02 (0.06)	0.74			
Language (English)	38	93	0.24			−23.62	0.00	0.95
versus Non-English				0.00 (0.08)	0.95			
Region (US)	38	93	0.23			−23.8	0.19	0.66
versus Non-U.S.				0.04 (0.07)	0.63			
Child age (Input)	38	93	0.24	0.00 (0.02)	0.95	−23.62	0.00	0.96
Child age (Assessment)	38	90	0.24	0.01 (0.02)	0.59	−20.78	0.15	0.70
Sources of input (Mother)	38	93	0.2			−22.43	5.68	0.22
versus Father				0.00 (0.07)	0.96			
versus Other				−0.09 (0.06)	0.32			
versus Primary				0.2* (0.03)	0.01			
versus All adults				0.1 (0.05)	0.06			
Mean input	33	83	0.22	−0.03 (0.02)	0.33	−18.97	0.97	0.32
Assessment characteristics								
Type (Direct)	38	93	0.3			−26.03	4.65	0.10
versus Report				−0.13* (0.06)	0.04			
versus Observed				−0.08 (0.07)	0.28			
Measure (Expressive)	38	93	0.21			−24.2	2.83	0.24
versus Receptive				0.09 (0.04)	0.07			
versus Both				0.05 (0.12)	0.73			
Vocabulary	38	93	0.24			−23.84	0.23	0.63
versus Non-Vocab				−0.02 (0.06)	0.70			
Study design								
Duration	31	82	0.23	0.01 (0.01)	0.16	−15.03	0.79	0.37
Context (Natural)	38	93	0.28			−21.72	2.63	0.45
versus Nat. Play				−0.12 (0.09)	0.23			
versus Struc. Play				−0.09 (0.07)	0.24			
versus Other				−0.05 (0.08)	0.58			
Location (Home)	38	93	0.23			−21.94	0.56	0.76

Table 2. *(Continued)*

Moderator	n	k	intcpt	β (SE)	p_β	AICc	χ^2	p_χ
versus Lab				0.03 (0.06)	0.65			
versus Other				−0.13* (0.03)	0.00			
Timeline (cross-lagged)	38	93	0.26			−24.73	1.12	0.29
versus Concurrent				−0.05 (0.03)	0.10			
Publication bias								
Peer-review status (PR)	38	93	0.21			−23.78	0.17	0.68
versus Non-PR				0.03 (0.08)	0.71			
Reported in PR Pubs.	38	93	0.23			−23.63	0.02	0.90
versus Not reported				0.01 (0.05)	0.88			

*p < 0.05; bolding indicates significant coefficient and χ^2.
intcpt = model intercept; AICc = Akaike information criterion (corrected); χ^2 = likelihood ratio test.

To check for moderating effects of publication type, we compared the non-peer-reviewed and studies with unreported statistics to our baseline. Neither of these variables were found to moderate the effect of word tokens on outcomes. In addition, we found no evidence of asymmetry in our funnel plot using Egger's test (β = 0.69, SE = 0.53, p = 0.22) (Figure 3). In sum, there was no evidence of publication bias in the word token studies.

3.2. Utterances

3.2.1. Summary statistics

We examined 45 correlations across 17 studies that measured word tokens (n = 956 unique participants). The number of effect size estimates per study ranged from 1 to 6 (median: 2). We found a medium-sized effect across studies (r = 0.19, Figure 4). Q-statistics revealed significant evidence for between-study heterogeneity (Q(44) = 338.75, p < 0.001). Some moderators were significant when included in the model (Table 3), but none of them improved overall model fit. Neither our analysis of publication status nor the Egger's test (β = 0.83, SE = 0.34, p = 0.08) revealed any evidence of publication bias (Figure 5).

3.3. Word types

We examined 111 correlations across 37 studies that measured word types (n = 2420 unique participants), with effect size estimates per study ranging from 1 to 10 (median: 3). We found a medium-sized effect across studies (r = 0.27, Figure 6). Q-statistics revealed significant evidence for between-study heterogeneity motivating an analysis of moderators (Q(110) = 621.82, p < 0.001). We found that studies with children who were older at the time of input data collection reported larger correlations (Table 4), significantly improving model fit (χ^2(1) = 12.64, p < 0.001). Studies with children who were older at the time of language outcome data collection also had marginally larger effect sizes (p = 0.05),

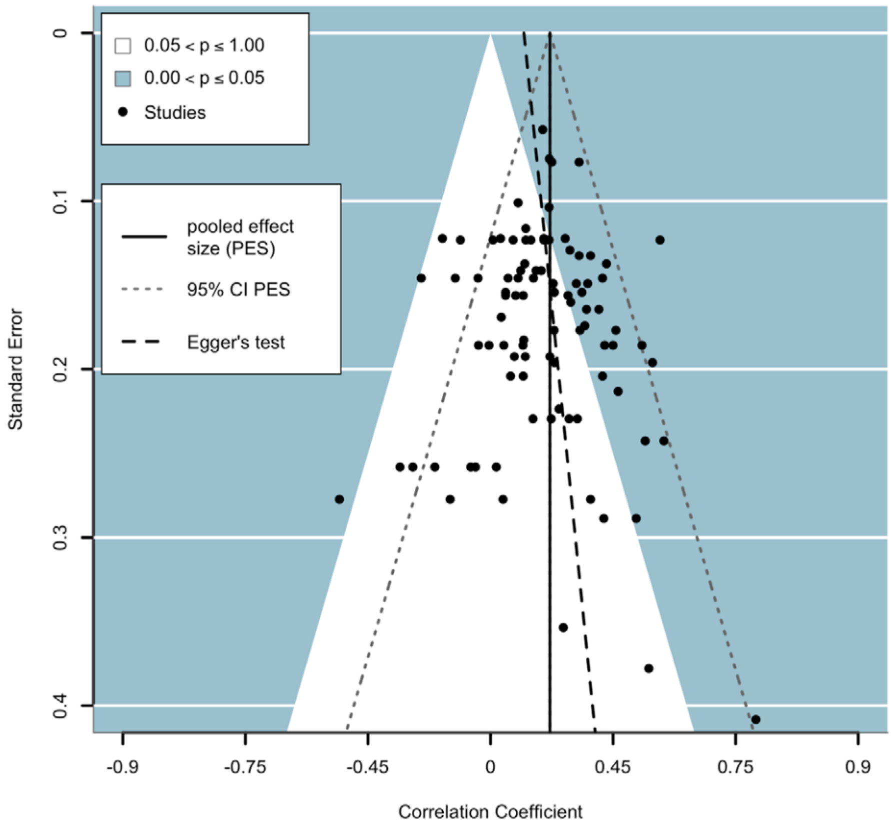

Figure 3. Funnel plot for studies of word tokens.

also improving model fit ($\chi^2(1) = 7.60, p = 0.006$). Unsurprisingly, we found that these two variables were highly correlated with one another ($R^2 = 0.39$, $p < 0.001$), and thus, it is unclear if it is the age at input, outcome, or both that predicts the bigger effect size.

We found that studies using parent-reported outcomes exhibited lower correlations with input than studies using direct assessments ($\chi^2(1) = 9.35, p = 0.03$). Parent reports are more commonly used when children are younger, resulting in a difference in child age across these assessment types ($\beta = -6.72$, $SE = 2.8$, $p = 0.02$). When assessment type is included in models with age at input collection and outcome assessment as predictors, it is non-significant and fails to improve model fit. Thus, this effect was most likely due to the partial confound with child age.

Finally, while including vocabulary matching (i.e., parent word types and child types produced during observation) in our model was found to improve fit, there were not significantly higher effect sizes in studies where input and outcome measures were matched (Table 3). As a follow-up, we included relation type as a main effect and interaction to this model to see whether this effect was significant for studies where input and outcome are collected during the same observation session (i.e., where parent input might situationally influence child output, or vice versa). We found a significant interaction between these variables, such that reported correlations were higher in concurrent

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

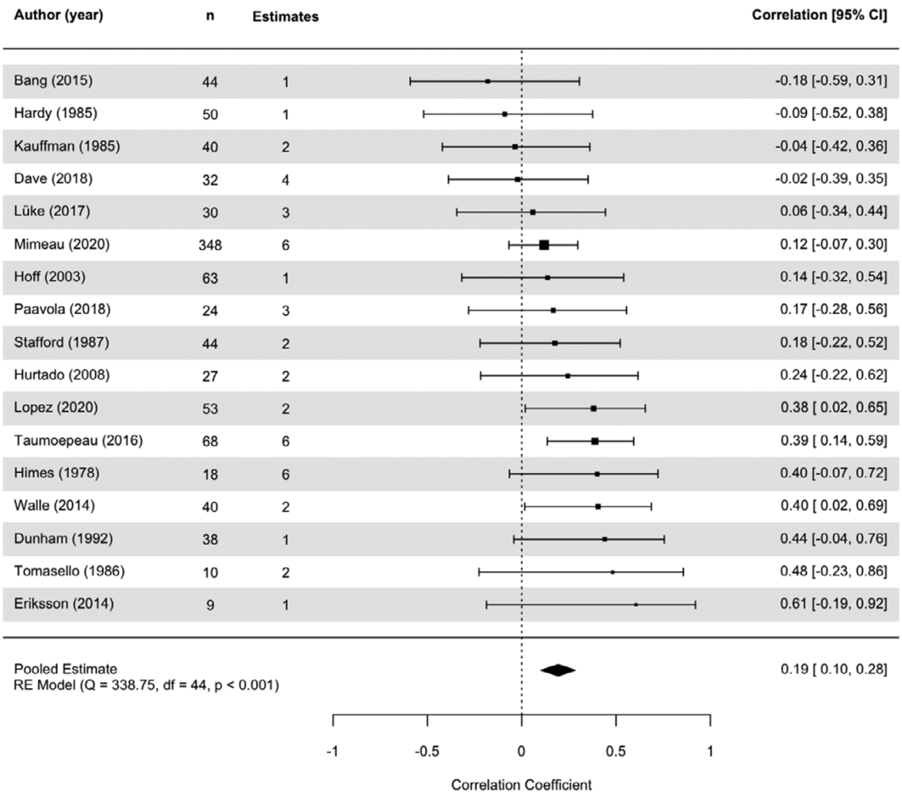

Figure 4. Forest plot of utterance study correlations.

Table 3. Summary of moderator analysis for studies of utterances

Moderator	n	k	intcpt	β (SE)	p_β	AICc	χ^2	p_χ
Base model	17	45	0.2			9.61		
Subject characteristics								
Child Gender (% Female)	17	45	0.22	0.09 (0.05)	0.22	9.67	2.35	0.13
Household SES (MU)	17	45	0.19			14.48	0.08	0.96
versus Diverse				0.02 (0.13)	0.89			
versus Low				0.06 (0.07)	0.36			
Language (English)	17	45	0.21			11.45	0.57	0.45
versus Non-English				−0.09 (0.11)	0.41			
Region (US)	17	45	0.18			11.84	0.19	0.67
versus Non-U.S.				0.04 (0.1)	0.70			
Child age (Input)	17	45	0.2	0.00 (0.07)	0.98	12.02	0.00	0.97

Table 3. *(Continued)*

Moderator	n	k	intcpt	β (SE)	p_β	AICc	χ^2	p_χ
Child age (Assessment)	17	45	0.2	−0.02 (0.04)	0.62	11.68	0.35	0.55
Sources of input (Mother)	17	45	0.2			8.64	5.92	0.05
versus Father				−0.47 (0.48)	0.50			
versus Other				0.23* (0.05)	0.00			
Mean input	13	38	0.18	−0.04 (0.06)	0.59	14.3	0.42	0.52
Assessment characteristics								
Type (Direct)	17	45	0.12			9.3	7.94	0.05
versus Report				0.07 (0.09)	0.49			
versus Observed				0.3 (0.23)	0.27			
versus Composite				−0.21* (0.04)	0.01			
Measure (Expressive)	17	45	0.22			12.87	1.69	0.43
versus Receptive				−0.04 (0.08)	0.61			
versus Both				−0.31* (0.07)	0.00			
Vocabulary	17	45	0.22			10.67	1.36	0.24
versus Non-Vocab				−0.12 (0.07)	0.15			
Study design								
Duration	17	45	0.33	0.47 (0.44)	0.39	10.65	1.37	0.24
Context (Natural)	17	45	0.33			13.58	3.66	0.30
versus Nat. Play				−0.16 (0.16)	0.41			
versus Struc. Play				−0.21 (0.12)	0.21			
versus Other				0.00 (0.15)	0.99			
Location (Home)	17	45	0.24			12.1	2.47	0.29
versus Lab				−0.04 (0.1)	0.71			
versus Other				−0.42* (0.08)	0.00			
Timeline (cross-lagged)	17	45	0.16			11.07	0.96	0.33
versus Concurrent				0.08 (0.08)	0.31			
Publication bias								
Peer-review status (PR)	17	45	0.22			12.01	0.02	0.89
versus Non-PR				−0.02 (0.23)	0.93			
Reported in PR Pubs.	17	45	0.28			10.32	1.7	0.19
versus Not Reported				−0.13 (0.1)	0.22			

*$p < 0.05$; bolding indicates significant coefficient and χ^2.
intcpt = model intercept; AICc = Akaike information criterion (corrected); χ^2 = likelihood ratio test.

Table 4. Summary of moderator analysis for studies of word types

Moderator	n	k	Intcpt	β (SE)	p_β	AICc	χ^2	p_χ
Base model	37	111	0.28			−19.13		
Subject characteristics								
Child Gender (% Female)	35	106	0.27	0.04 (0.04)	0.41	−18.11	1.14	0.28
Household SES (MU)	37	111	0.25			−16.09	1.97	0.37
versus Diverse				0.13 (0.11)	0.27			
versus Low				0.02 (0.06)	0.71			
Language (English)	37	111	0.28			−16.82	0.52	0.47
versus Non-English				−0.07 (0.06)	0.31			
Region (US)	37	111	0.29			−17.03	0.72	0.39
versus Non-U.S.				−0.07 (0.09)	0.45			
Child age (input)	**37**	**111**	**0.26**	**0.1* (0.04)**	**0.02**	**−28.95**	**12.65***	**0.00**
Child age (assessment)	37	108	0.25	0.08 (0.04)	0.05	−21.94	7.6*	0.01
Sources of input (Mother)	37	111	0.28			−14.67	5.08	0.28
versus Father				−0.11 (0.07)	0.15			
versus Other				−0.06 (0.14)	0.72			
versus Primary				0.15 (0.11)	0.38			
versus All Adults				0.12* (0.04)	0.01			
Assessment characteristics								
Type (Direct)	**37**	**111**	**0.29**			**−21.23**	**9.35***	**0.02**
versus Report				**−0.14* (0.06)**	**0.03**			
versus Observed				0.03 (0.07)	0.64			
versus Composite				−0.08 (0.04)	0.06			
Measure (Expressive)	37	111	0.28			−14.47	0.36	0.84
versus Receptive				−0.03 (0.07)	0.65			
versus Both				0.01 (0.05)	0.90			
Vocabulary	37	111	0.29			−17.3	1.00	0.32
versus Non-Vocab				−0.05 (0.03)	0.15			
Matched input/Outcome	37	111	0.38			−28.22	11.91*	0.00
versus Non-Matched				−0.15 (0.07)	0.06			
Study design								
Duration	36	110	0.26	−0.01 (0.24)	0.98	−23.06	0.00	0.96
Context (Natural)	37	111	0.33			−14.16	2.28	0.52
versus Nat. Play				−0.08 (0.12)	0.52			
versus Struc. Play				−0.1 (0.11)	0.34			
versus Other				−0.02 (0.13)	0.88			
Location (Home)	37	111	0.28			−14.27	0.15	0.93

Table 4. *(Continued)*

Moderator	*n*	*k*	Intcpt	β (SE)	p_β	AICc	χ^2	p_χ
versus Lab				−0.03 (0.07)	0.71			
versus Other				0.02 (0.05)	0.64			
Timeline (cross-lagged)	37	111	0.25			−17.25	0.94	0.33
versus Concurrent				0.04 (0.05)	0.39			
Publication bias								
Peer-Review status (PR)	37	111	0.3			−16.46	0.15	0.70
versus Non-PR				−0.03 (0.11)	0.77			
Reported in PR Pubs.	37	111	0.27			−16.33	0.02	0.89
versus Not reported				0.01 (0.07)	0.90			

*$p < 0.05$; bolding indicates significant coefficient and χ^2.
intcpt = model intercept; AICc = Akaike information criterion (corrected); χ^2 = likelihood ratio test

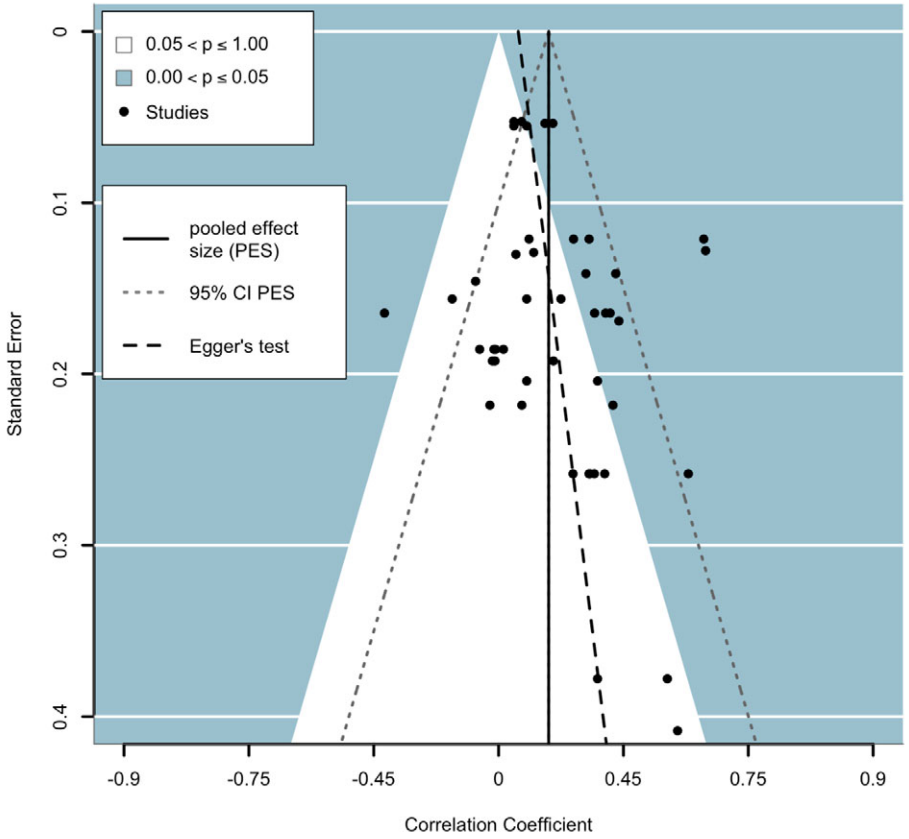

Figure 5. Funnel plot for studies of utterances.

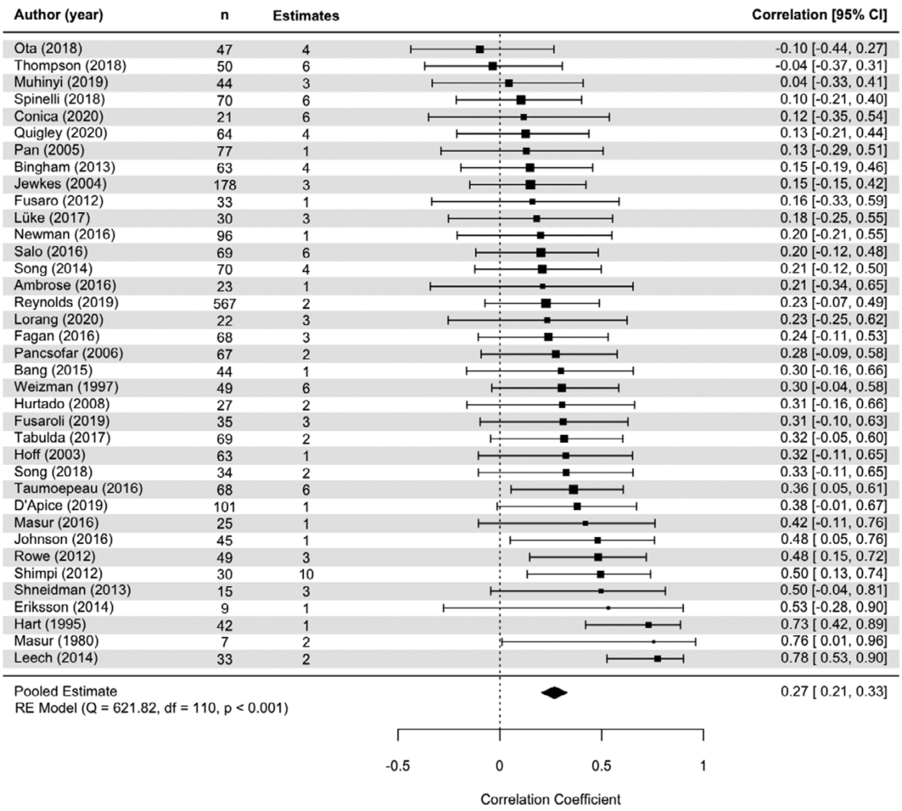

Figure 6. Forest plot of word type study correlations.

studies where input and outcomes were measured in the same way (main effect of matching: $\beta = -0.01$, $SE = 0.08$, $p = 0.88$; interaction: $\beta = 0.23$, $SE = 0.07$, $p = 0.01$).

In our analysis of publication bias, neither peer review nor whether the correlation was reported was found to moderate the effect of word types on outcomes. However, we did find evidence of funnel plot asymmetry using Egger's test ($\beta = 1.22$, $SE = 0.41$, $p = 0.02$), such that studies with larger standard errors tended to be skewed towards larger positive values (Figure 7).

3.4. MLU

We examined 74 correlations across 27 studies that measured mean length of utterance ($n = 2340$ unique participants), with the number of effect size estimates per study ranging from 1 to 6 (median: 3). We found a medium-sized effect across studies ($r = 0.21$, Figure 8). Q-statistics revealed significant evidence for between-study heterogeneity (Q (74) = 339.42, $p < 0.001$), motivating an analysis of possible moderators. We found a significant and positive correlation between the length of the observation session and effect size (Table 5, $\chi^2(1) = 12.92$, $p < 0.001$), with longer studies producing larger effect sizes. We might expect to see such an effect if MLU measures were more stable when the

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

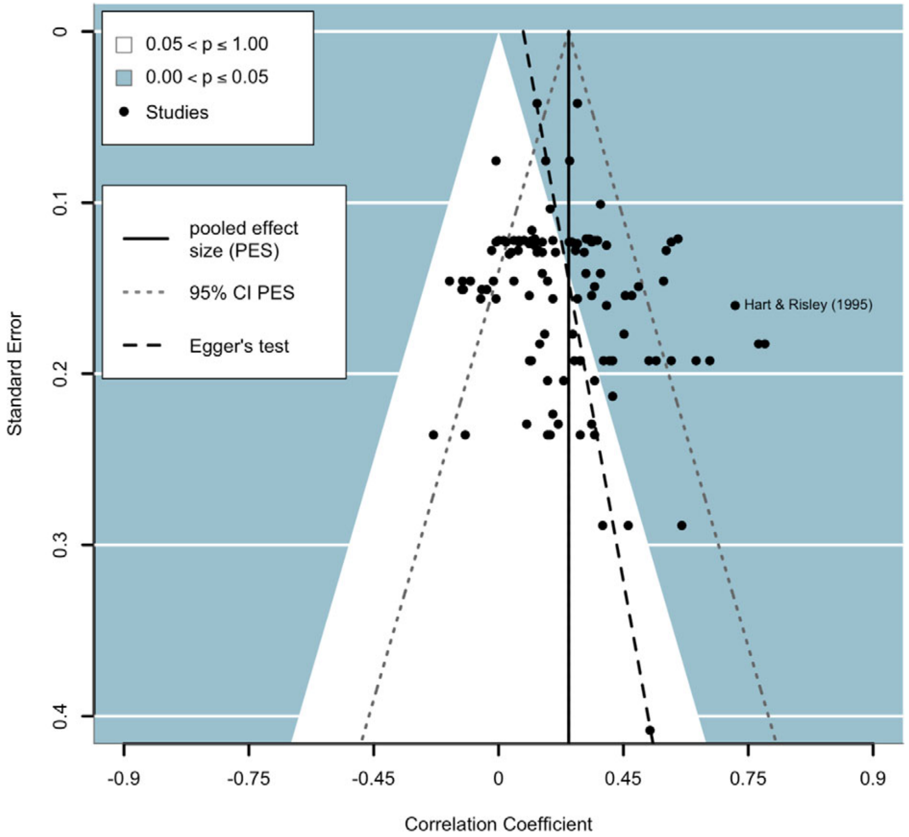

Figure 7. Funnel plot for studies of word types. Hart and Risley (1995) is given for comparison.

sample of utterances is larger. Neither our analysis of publication status nor the Egger's test ($\beta = 0.54$, $SE = 0.47$, $p = 0.25$) revealed any evidence of publication bias (Figure 9).

3.5. Comparison of input measures

Finally, to compare the effect sizes for our four input measures, we first constructed a baseline model with no moderators, containing 323 input–outcome correlations across all 71 studies (Table 6). Overall, there was a medium-sized association between all input and children's language outcomes ($r = 0.24$, $p < 0.001$; CI [0.20; 0.29]). Next, we added input measure as a moderating variable, using tokens as the contrast case. No significant difference in effect size was found between our input measures, and there was no improvement of model fit.

4. Discussion

The present meta-analysis drew upon 71 studies and 4760 participants to explore the magnitude of input effects. The analysis included 38 studies that were not included in the

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

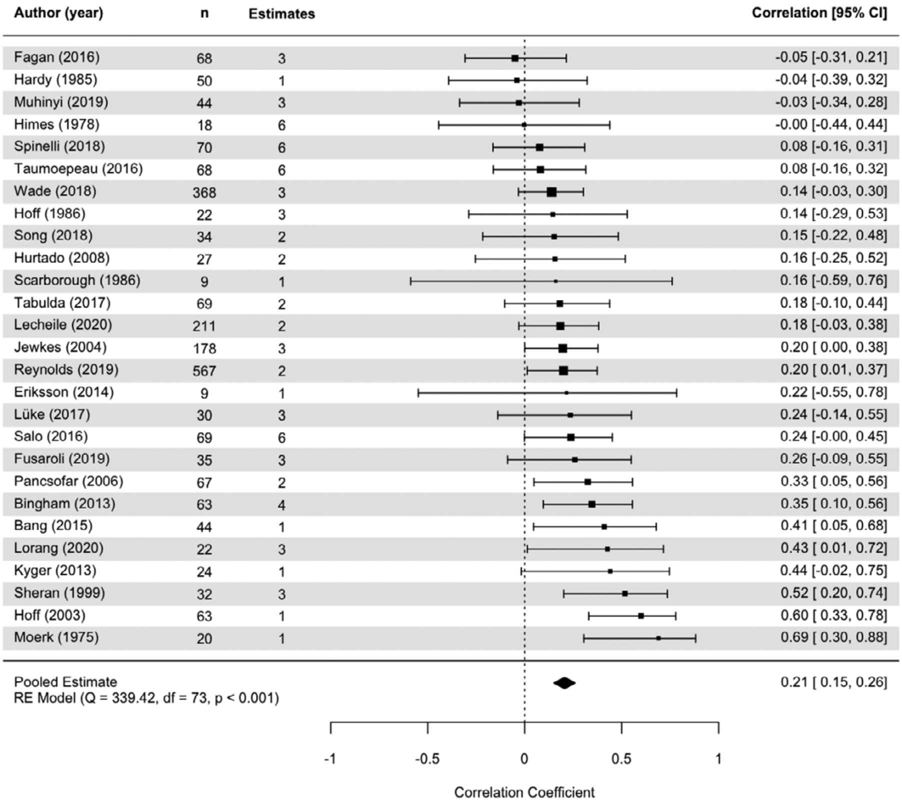

Figure 8. Forest plot of MLU study correlations.

most recent prior meta-analysis on this topic (AGPJM). In addition, our analysis employed an innovative statistical method that allowed us to include multiple effect sizes per study, resulting in a total of 323 effect sizes. As yet there are no widely accepted methods for determining power in multilevel meta-regression models (e.g., Vembye et al., 2023). Nevertheless, we should expect, on first principles, that the accuracy and sensitivity of an analysis will increase as the number of studies and effects that are included increases. Our meta-analysis also contributed to this literature by including studies conducted on languages other than English and using more sensitive within-study comparisons to explore differences in effect size across measures.

We found that the relationship between caregiver input and child language outcomes is reliable across four different measures of caregiver input: utterances, word tokens, word types, and MLU. These measures all produced similar small-to-medium-sized effects, with no significant differences between them. For word types, we found that effect sizes were reliably larger when children were older. For MLU, we found that effect sizes were larger in studies with longer observation sessions. We also found evidence that using parent and child word types collected from the same session produce larger correlations.

However, most of the moderators that have been hypothesized to be relevant were not reliable predictors of effect size in our analyses. This included caregiver demographics,

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

Table 5. Summary of moderator analysis for studies of MLU

Moderator	*n*	*k*	intcpt	β (SE)	p_β	AICc	χ^2	p_χ
Base model	27	74	0.21			−30.01		
Subject characteristics								
Child Gender (% Female)	25	68	0.22	−0.02 (0.02)	0.31	−28.26	0.51	0.47
Household SES (MU)	27	74	0.24			−32.82	1.8	0.41
versus Diverse				−0.04 (0.07)	0.54			
versus Low				−0.12 (0.12)	0.40			
Language (English)	27	74	0.21			−33.4	0.08	0.78
versus Non-English				−0.02 (0.09)	0.80			
Region (US)	27	74	0.23			−34.96	1.64	0.20
versus Non-U.S.				−0.08 (0.05)	0.15			
Child age (input)	27	74	0.21	−0.01 (0.05)	0.84	−33.43	0.1	0.75
Child age (assessment)	27	74	0.21	−0.03 (0.04)	0.50	−34.2	0.87	0.35
Sources of input (Mother)	27	74	0.21			−31.4	0.38	0.83
versus Father				−0.03 (0.06)	0.57			
versus Other				0.05 (0.03)	0.15			
Mean input	23	68	0.2	−0.01 (0.05)	0.89	−38.84	0.04	0.84
Assessment characteristics								
Type (Direct)	27	74	0.21			−32.49	3.84	0.28
versus Report				−0.04 (0.05)	0.41			
versus Observed				0.03 (0.04)	0.50			
versus Composite				−0.25* (0.03)	0.00			
Measure (Expressive)	27	74	0.22			−32.1	1.08	0.58
versus Receptive				−0.05 (0.04)	0.26			
versus Both				0.00 (0.07)	0.99			
Vocabulary	27	74	0.18			−36.86	3.53	0.06
versus Non-Vocab				0.08 (0.04)	0.10			
Matched input/Outcome	27	74	0.25			−33.76	0.44	0.51
versus Non-Matched				−0.04 (0.07)	0.59			
Study design								
Duration	**26**	**73**	**0.4**	**0.73* (0.2)**	**0.01**	**−46.38**	**12.92***	**0.00**
Context (Natural)	27	74	0.34			−33.48	4.83	0.18
versus Nat. Play				−0.15 (0.26)	0.60			
versus Struc. Play				−0.11 (0.26)	0.72			

Table 5. *(Continued)*

Moderator	*n*	*k*	intcpt	*β* (SE)	p_β	AICc	χ^2	p_χ
versus Other				−0.2 (0.26)	0.51			
Location (Home)	27	74	0.25			−36	4.98	0.08
versus Lab				−0.11 (0.06)	0.07			
versus Other				0.19* (0.04)	0.00			
Timeline (cross-lagged)	27	74	0.18			−35.01	1.69	0.19
versus Concurrent				0.07 (0.06)	0.26			
Publication bias								
Peer-Review status (PR)	27	74	0.25			−33.78	0.46	0.50
versus Non-PR				−0.05 (0.07)	0.53			
Reported in PR Pubs.	27	74	0.26			−34.69	1.37	0.24
versus Not Reported				−0.07 (0.07)	0.33			

*$p < 0.05$; bolding indicates significant coefficient and χ^2.
intcpt = model intercept; AICc = Akaike information criterion (corrected); χ^2 = likelihood ratio test.

child demographics, and whether the measure was based on a speech sample, experimenter administered test, or parent report. Critically, we did not replicate three findings from the AGPJM meta-analysis. In our sample, we found no evidence that naturalistic studies have larger effect sizes than more structured observations, nor that studies with cross-lagged observations have larger effects than studies with concurrent observations. Furthermore, we found no evidence to support the claim that measures of input quality are more reliable predictors than measures of input quantity, despite using potentially more sensitive within study models.

Finally, we found evidence for publication bias for studies of where parental word types were the critical input variable. In contrast, AGPJM found evidence for publication bias in studies of input quantity (i.e., tokens and utterances).

The remainder of our discussion we address four issues: (1) assessing the pooled effect sizes observed in this meta-analysis and how it affects our understanding of the input literature and its policy implications; (2) interpreting the moderators observed in our analysis; (3) understanding the null effects in this analysis; and (4) the limitations of this meta-analysis and the input literature more broadly. Throughout our discussion, we will conduct exploratory analysis on critical subgroups of studies within our sample to rule out different hypotheses about our results.

4.1. Assessing the magnitude of the pooled effects

A central goal of meta-analysis is to better understand how large a particular effect truly is. In our analyses, we found pooled effect sizes that ranged from $r = 0.19$ to $r = 0.27$. It is easier to conceptualize these effects if we convert them to R^2 so that they represent the proportion of variance accounted for by the input variable. On this scale, the effects range from $R^2 = 0.04$ for utterances to $R^2 = 0.07$ for types. These estimates are quite similar to those in AGPJM even though less than half (46%) of the studies in our sample appeared in

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

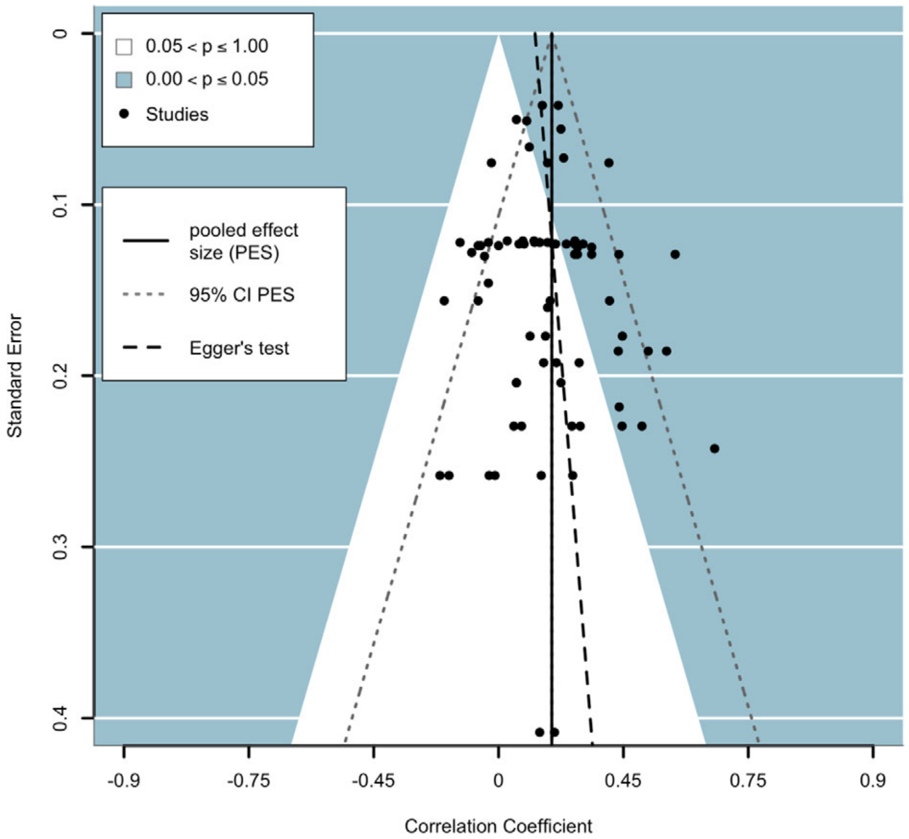

Figure 9. Funnel plot for studies of MLU.

Table 6. Comparison of pooled effect sizes across input measures

Moderator	*n*	*k*	intcpt	*β* (SE)	p_β	AICc	χ^2	p_χ
Base model	71	328	0.25			−58.76		
Input measure (Word Types) versus Word Tokens	71	328	0.27	−0.03 (−0.03)	0.17	−55.88	5.39	0.25
versus MLU				−0.07 (−0.07)	0.14			
versus Utterances				−0.02 (−0.02)	0.77			
versus Composite				0.03 (0.03)	0.89			

*p < 0.05; bolding indicates significant coefficient and χ^2.
intcpt = model intercept; AICc = Akaike information criterion (corrected); χ^2 = likelihood ratio test.

their sample and the outcome measures were categorized differently. Specifically, in AGPJM, the estimates ranged from $R^2 = 0.04$ for quantity to $R^2 = 0.11$ speech complexity.

These estimates might seem modest to those who were introduced to this question by Hart & Risley's seminal 1995 study. It is hard to overstate the effect that H&R have had on

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

language acquisition research and social policy; as of July 2024, Google Scholar lists over 12,000 citations to their 1995 book. However, the magnitude of the input–outcome relationship found by H&R for caregiver word types is substantially higher than the pooled effect size found in our meta-analysis ($R^2 = 0.53$ versus $R^2 = 0.07$). This difference would be critical, for example, for our expectations about the impact of a policy that sought to improve child language outcomes with parent training. Thus, understanding this discrepancy is critical to understanding how we can best use limited resources.

There are two broad explanations for these divergent effect size estimates. First, the H&R study might have unique properties that lead the true effect to be larger in their sample. Second, it is possible that H&R is simply one sample drawn from an underlying distribution in which the true effect size is roughly equivalent to the estimate from our analysis. The first explanation could lead to new directions for research and new ways in which policies might be targeted. The second hypothesis suggests that we, as scientists and policy makers, may need to adjust our expectations.

There are several features of H&R that stand out as potential reasons for a larger effect size. First, their measures of parent speech and their outcome measures were based on large samples of speech, collected over an unusually long-time frame, in a naturalistic context. Specifically, as many as 29 hour-long observation sessions were conducted in the child's home between the ages of 7 months to 3 years. In contrast, the other studies of word types in our sample used between 2 minutes and 2 hours of input ($M = 26$ minutes). As a result, H&R may have produced more accurate estimates of parental speech resulting in a larger observed correlation. Similarly, their primary outcome variable was an estimate of lexical types produced across three-hour-long observation sessions. The mean length of the child observations across studies of word types in our sample was 15 minutes. If these factors were responsible for the larger effect size, it would suggest (1) that policymakers could expect large effects from interventions that are effective in changing parental input in enduring ways, and (2) that researchers (and clinicians) should consider longer data collection periods for input studies. Our meta-analysis, however, does not support this conclusion: for studies that used parental types as an input measure, we found no evidence that the length of the observation period moderated the effect size. One might question the relevance of our moderator analysis, since there are few studies with an observation period that was anywhere near as long as H&R. We disagree: if the advantage of larger speech samples is that they are less noisy, then we would expect to see the steepest improvement in stability at the low end of the scale. This is, however, ultimately an empirical question. The relationship between sample size and correlation strength could be directly tested by conducting secondary analyses of the H&R data set to determine how rapidly input measures approximate the estimate from the total sample as increasingly large subsamples of input are analysed.

The second feature that makes H&R unusual is that the sample was selected to overrepresent the extreme ends of the socio-economic spectrum in the U.S. Of the 42 participating families, six were receiving welfare benefits and thirteen were recruited because the primary wage earner was a high-status professional. In their sample, there was a strong relationship between SES and input, with caregivers in professional families producing about three times as much speech as the parents receiving benefits. Subsequent studies have not found differences between their SES groups that are anywhere near this large (see Dailey & Bergelson, 2022). For example, Hoff (2003) found that the high SES families in her sample produced roughly 33% more speech than her mid SES group (see also Gilkerson et al., 2017). It is unclear whether this difference in findings reflects the unusual composition of the Hart and Risley sample, changes in child-rearing practices across communities in the U.S. over time, or the way in which their input measures were collected and conceptualized. But critically, whatever its cause, the tighter link between SES and input in the Hart and Risley sample raises the possibility that the unusually large correlation between input and outcome in that study is attributable to other causal pathways linking parental SES to child language outcomes, such as passive gene environment correlations (see Coffey et al., 2022 for discussion) or the effect of higher maternal education on a wider range of parenting practices that might influence children's linguistic and cognitive development.

As we noted above, the second hypothesis is that H&R is drawn from the same underlying distribution as the other studies with a true effect size around $R^2 = 0.07$. H&R's sample size was modest for a study focused on individual differences ($n = 42$). We expect the variability of effect sizes to increase as sample size decreases (giving funnel plots their characteristic shape). The H&R results, however, fall outside of the range of what we might expect on that basis alone (see Figure 7). In fact, even if the true effect size was moderately larger than our estimate (e.g., $r = 0.35$, or the upper limit of our 95% confidence interval), H&R would remain an outlier.

4.2. Moderator effects

4.2.1. Older children benefit more from lexical diversity

We found that the pooled effect size for word type studies was larger when the children studied were older at both the time of input collection and the time of language assessment. Why might younger children benefit less from lexical diversity?

One possibility is that words that children acquire early in life are so common and so concrete (Braginsky et al., 2019; Coffey & Snedeker, 2024) that they are likely to appear in informative contexts even in the speech of parents who exhibit lower lexical diversity. After children pick up these more common words, they learn words that are less frequent and less consistent across parents but are still quite concrete, like *tiger* or *truck* (Coffey et al., 2024). At this stage, children who hear more diverse input could reap the benefit of encountering more word types. In addition, as children become more linguistically proficient, they are more likely to learn words that are frequent but not particularly concrete (Coffey et al., 2024). Words of this kind can often only be acquired by using information from different contexts in which the word was used (Gillette et al., 1999). More lexically diverse speech might be more likely to provide these clues.

We did not find any evidence that age moderated the effects of the other input measures. The fact that the quantity of input (tokens and utterances) is equally helpful across this developmental range is consistent with most learning theories—more learning opportunities are helpful even for the easiest words. However, to the extent that these measures are correlated with types (see below), we would expect to find an effect of age given a sufficiently large sample of studies and participants.

The prior literature on MLU has been mixed. Some conclude that it only predicts outcomes when it is tailored to children's level of development (in which case we *would* expect an effect of age, e.g., Murray et al., 1990). Others find that complex speech predicts outcomes at all ages (in which case we would *not* expect an effect of age, e.g., Hoff-Ginsberg, 1986). The fact that MLU *does* relate to input across studies but is *not* moderated by age suggests the latter may be true.

Interestingly, while AGPJM found an effect of child age on correlations with input quality, when they conducted separate analyses on studies that contained measures of lexical diversity (equivalent to our word types) and sentence complexity (consisting of not only MLU, but also other measures, such as sophistication, rare words, and multi-clausal utterances), they found no effects. Our results suggest that the age effects in the primary analysis were likely driven by word types, rather than the complexity. Our ability to find this effect within the studies of lexical diversity is likely due to the larger sample of relevant studies available to us (N = 17 versus N = 37).

4.2.2. Length of the observation and MLU

We found that MLU studies with longer observation sessions reported larger correlations. This could be because longer observation sessions produce more stable measures of MLU. However, this leaves open the question of why we did not find a moderating effect in our analyses of utterances, word tokens, or word types. One possibility is that MLU is intrinsically a less stable measure than the others, requiring a longer session to measure reliably. This could be assessed using existing data sets (e.g., by comparing correlations across different-sized sub-samples). AGPJM also found that observation duration moderated the effect size of input quality studies. Our results suggest that this finding may be driven by studies using sentence complexity measures, rather than lexical diversity measures.

4.3. Surprising non-moderators

4.3.1. Observation activity

We did not find significant differences in the size of the input–outcome correlations depending on the activity during the observation session. In contrast, in their analysis of input quality, AGPJM found that studies using naturalistic observation produced larger effect sizes, as compared to other contexts. A priori, we might expect larger effect sizes from naturalistic observations because they might be more representative of typical input. Previous studies, however, have found that input measures from structured and naturalistic observations are correlated (Tamis-LeMonda et al., 2017).

One possibility is that we reduced the difference in effect size between naturalistic studies and other studies by including LENA studies, which are naturalistic but were omitted by AGPJM. We do not believe that this is the case: omitting LENA studies from our analysis did not change our results (see Supplementary Materials). We do not find this surprising, as meta-analyses of LENA studies and non-LENA studies result in similar effect-size estimates (Wang et al., 2020; AGPJM).

Instead, we suspect that the finding in AGPJM is attributable to the very small number of naturalistic studies in their sample (5 for the quality analysis). This could make their moderator analysis vulnerable to skewing due to a couple of naturalistic studies with unusually large effect sizes (such as H&R). In contrast, our sample of naturalistic studies was larger (10 for types), which may have made our analysis less sensitive to skewing.

4.3.2. Cross-lagged versus concurrent studies

We found no significant differences between studies where input and outcome are collected concurrently and studies where data collection was cross-lagged. In contrast, AGPJM found larger correlations in quantity studies that were cross-lagged. This is

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

unlikely to reflect skew due to outliers since there are a number of studies of both kinds in their sample ($N = 16$ for concurrent, $N = 17$ for cross-lagged).

Given our large sample of effect sizes ($k = 93$ for tokens; $k = 45$ for utterances), it is unlikely that we lacked the power to detect such an effect. Instead, we suspect that their finding was a side-effect of their hierarchical data selection procedure: if studies had cross-lagged and concurrent correlations, only a cross-lagged correlation was included in the meta-analysis, creating a confound between study complexity/length and temporal design. In our study, all correlations were included. In addition, given the large number of mediators in these meta-analyses and the confounds between them, we are likely to find effects that shrink or disappear as more data are collected (Barnett et al., 2005; Gelman & Carlin, 2014).

4.3.3. Quality versus quantity of input

Another notable difference between our study and AGPJM is the fact that we did not find a difference in effect size between any of our input measures. Some researchers have argued that measures of input *quality* are better suited to predict individual differences in language outcomes than measures of input *quantity* (e.g., Golinkoff et al., 2019). This would be expected if the pace of acquisition did not depend primarily on the number of words a child encounters but the degree to which the context of word use allows them to infer their meaning. While the logic behind this argument is sound, one might still expect the effect sizes for quantity and quality measures to be quite similar because in practice they are often highly correlated. To explore this, we calculated these correlations for the studies in our sample with available data. There were large correlations between types and tokens (range: $r = 0.65$–0.94; median: $r = 0.88$; $k = 9$), types and utterances (range: $r = 0.45$–0.90; median: $r = 0.75$; $k = 6$), and MLU and tokens (range: $r = 0.19$–0.71; median: $r = 0.44$; $k = 8$). The only correlation that was small and sometimes negative was between utterances and MLU (range: $r = -0.44$–0.36; median: $r = 0.12$; $k = 8$).

4.4. Remaining questions

4.4.1. Culture and language

The dearth of input studies conducted outside of the Western world or with speakers of non-Western languages makes a systematic investigation of cultural or linguistic moderators of the input–outcome relationship difficult. In our analysis, we tried to get at this question by characterizing studies as either “within the U.S.” or “outside the U.S.” and as either “English” or “non-English.” This classification system cannot be justified on cultural, geographic, or linguistic grounds. It makes sense only in light of the degree to which developmental research in general, and work on this topic in particular, has focused on English-speaking populations within the United States (Kidd & Garcia, 2022). This was the only coding scheme that would allow us to amass a reasonable, albeit small, number of studies in the second group.

Nevertheless, we see this as one small but important step in using meta-analytic approaches to examine cross-cultural input studies. We found no evidence that studies conducted in English or in the U.S. produced larger or smaller correlations than other studies. This finding has two critical limitations. First, due to the small number of non-English studies (12/75) and non-US studies (16/75), we may lack the power to detect modest effects. Second, the non-English and non-US samples consisted of families living

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

in urban areas of Europe, East Asia, or North America. While there are a few input studies conducted in rural agrarian settings (e.g., Mastin & Vogt, 2016; Shneidman & Goldin-Meadow, 2012; Weber et al., 2017; Zhang et al., 2023), these studies were not eligible for our meta-analysis for a number of reasons (e.g., no appropriate input measures, no input–outcome correlations reported, or conducted after the final search). Thus, we cannot speak to the degree to which input–output correlations vary across the full range of human societies.

Cross-cultural research is critical for understanding the nature of input–outcome correlations and what they might reveal about the causal role of input in language development. There is considerable cross-cultural variation in how parents speak to their children and their beliefs about the role this plays in language development (Schieffelin & Ochs, 1986). Our current understanding of the relationship between input variation and outcome is based almost entirely on a narrow set of environments (mostly in the U.S., mostly in English) in which talking to young children is not only accepted but encouraged and deemed valuable. Determining whether the magnitude of these input–outcome correlations is affected by variation in mean input amount or variation in the language socialization practices will provide critical insights into the causal connections between input and outcome. If the correlations with caregiver speech shrink or disappear entirely in some contexts, it might suggest that other sources of input play a larger role in these contexts, that additional factors need to be present for input to set the pace for outcomes, or that third variables (like maternal education) are inflating the correlations in WEIRD societies. If the correlations are present cross-culturally but increase with variation in input within the population, it would provide additional support for the simple causal model in which input sets the pace for early acquisition.

Currently, however, we are in no position to address these questions. While our analysis of mean input as a moderator was negative for all input measures, the range of variation was restricted to what is found in urbanized societies where formal education is valued. Our review confirms the need for additional research on input effects in non-Western societies, small-scale societies, agrarian societies, and societies where secondary education is less common.

4.4.2. Socioeconomic status

There are compelling reasons to believe *a priori* that we would find larger input effects in studies of low-SES households. Environmental differences account for more variance in the developmental outcomes of children from lower-SES households than children from higher-SES households (Turkheimer et al., 2003). One explanation for this pattern is that children in lower-SES households experience more environmental heterogeneity than high-SES children. If this was the case, we might expect to see larger input–outcome correlations in studies with low-SES households, which we did not. One possibility is that we were underpowered to find such effects. We had fewer studies that drew only from low-SES households as compared to middle-upper SES households (e.g., $N = 7$ Low versus $N = 24$ M-U in our analysis of word types).

We might have also expected to find differences larger correlations in studies that contain socioeconomically diverse samples of children (relative to middle-upper SES samples). Recent studies have confirmed that there are, on average, modest but reliable differences in the amount of child-directed speech between higher-SES and lower-SES households (Dailey & Bergelson, 2022). Thus, we would expect studies that sampled from different socioeconomic groups would find greater variation in input, and therefore larger

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

input–outcome correlations. However, the absence of a moderating effect in both AGPJM's meta-analysis and our own suggests that this is not the case. Here again, power is a concern: there are fewer studies in our analyses that draw from different socio-economic groups (e.g., $N = 6$ Diverse versus $N = 24$ M-U in our analysis of word types).

4.5. Limitations

4.5.1. Unable to establish causality

Although our approach has demonstrated that associations between input and outcome are reliable across studies, correlational research of this kind cannot disambiguate the causal relationship between these factors. It is possible that the robust relationships between input and outcome we observe are caused by a third variable that influences both parental speech and the pace of child language acquisition. For example, input could be related to other environmental factors that impact development, such as general parental attentiveness or the availability of educational materials in the home. In addition, most input studies are conducted with children and their biological parents. Language ability, like most human characteristics, is greatly influenced by genetic factors (Polderman et al., 2015; Stromswold, 2001). This introduces the possibility that the associations between caregiver input and children's language outcomes we observe are genetic in nature: verbal parents have verbal children because they pass on those genes.

Nevertheless, there are several reasons to believe that these effects might be causal in nature. For one, parent-targeted randomized control trials that produce changes in input also often impact language outcomes (e.g., Suskind et al., 2016; Weber et al., 2017). Second, as we have seen in our meta-analysis, input effects persist across a range of environments and measures, suggesting that, if there is a third variable underlying the pattern, it must be one that is correlated with both input and outcome across these environments. Finally, although there are only a few studies that use genetically non-confounded designs, these studies find reliable input–outcome correlations (Hardy-Brown et al., 1981; Huttenlocher et al., 2002; Gauthier et al., 2013; Coffey et al., 2022, but see Wadsworth et al., 2002). Unfortunately, there are not enough studies of this kind to use meta-analysis to determine whether the input–outcome correlation in these studies is smaller than studies with a genetic confound. Future work of this kind is necessary to understand the complex causal pathways linking language input and outcomes.

4.5.2. Remaining sources of bias

The only evidence of publication bias that we found was asymmetry in the funnel plot for word types, indicating that studies with smaller samples reported larger effects than we would expect. This asymmetry could reflect differences in the methods used in larger and smaller studies, but it could also result from studies with non-significant correlations being culled from the literature. We attempted to address this by reaching out to authors for unpublished studies, but this approach is unlikely to totally eliminate this source of bias. In a paper examining 10 meta-analyses across different areas of language and cognitive development, Tsuji et al. (2020) found that the inclusion of unpublished literature did not result in any significant difference in estimated effect size. This may be because unpublished data is often collected by reaching out to authors in familiar networks which may favour the reporting of positive results. Furthermore, it is likely that in many published studies only a subset of correlations calculated between variables were

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

reported. Preregistration and open data access have emerged as partial solutions to this problem.

4.5.3. Limited data on other kinds of input

Almost all of the studies considered here tracked child-directed speech produced by adults. An open question in language development is the degree to which children benefit from overheard speech or speech produced by other children. Many accounts of rural societies stress the importance placed on children's ability to learn about the adult world by watching and listening (e.g., Schieffelin & Ochs, 1986; Shneidman & Goldin-Meadow, 2012). In many cultures, older children assume caregiving responsibilities early in life and potentially account for a large amount of input to young learners (e.g., Loukatou et al., 2022; Shneidman & Goldin-Meadow, 2012). Some previous studies have suggested that overheard speech and sibling speech are less useful for learners, at least in WEIRD societies (e.g., Mannle et al., 1992). This might lead us to expect smaller or non-significant relationships with outcomes as compared to maternal input. Within our sample, the few studies of overheard speech ($N = 3$) and sibling input ($N = 2$) give uniformly null results. Nevertheless, omitting these sources of speech risks mischaracterizing the early language environments of children in other cultural contexts (Sperry et al., 2019).

4.6. Conclusion

In our sample of 71 input studies, we found that caregiver input predicted child language outcomes, albeit to a lesser degree than some early studies suggested ($R^2 = 0.04$–0.07). The size of these input–outcome associations is similar across different input measures. For word types, we found evidence that the correlation increases with age, as well as evidence of publication bias. For mean length of utterance, we found larger associations between input and outcome measures in longer observation sessions.

Supplementary material. The supplementary material for this article can be found at http://doi.org/10.1017/S0305000924000692.

Data availability statement. Our data are fully available via OSF: OSF | Does Talking To Children Matter? A Meta-Analysis (https://osf.io/aydcf/).

Acknowledgements. We would like to thank all the authors who contributed publications and data to this project: Carina Lüke, Meredith Rowe, Elaine Smolen, Ronda Rufsvold, Maria Hartman, Jill Gilkerson, Erika Hoff, Riccardo Fusaroli, Eric Walle, Laurie Hoffman, Mitsuhiko Ota, Mele Taumoepeau, Katrina D'Apice, Gabrielle Strauss, Naja Ferjan Ramírez, Daniel Swingley, Shuxia Liu, Aleka Akoyunoglou Blackwell, Allegra Cattani, Katie Alcock, and Christine Cox Eriksson. We would also like to thank our research assistant Claire Lin for her work on data collection and coding.

Competing interest. We have no known competing interests to disclose.

References

Anderson, N. J., Graham, S. A., Prime, H., Jenkins, J. M., & **Madigan, S.** (2021). Linking quality and quantity of parental linguistic input to child language skills: A meta-analysis. *Child Development*, **92**(2), 484–501. https://doi.org/10.1111/cdev.13508

Barnett, A. G., Van Der Pols, J. C., & **Dobson, A. J.** (2005). Regression to the mean: What it is and how to deal with it. *International Journal of Epidemiology*, **34**(1), 215–220. https://doi.org/10.1093/ije/dyh299

Bates, D., **Mächler, M.**, **Bolker, B.**, & **Walker, S.** (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, **67**(1), 1–48.doi:https://doi.org/10.18637/jss.v067.i01

Braginsky, M., **Yurovsky, D.**, **Marchman, V. A.**, & **Frank, M. C.** (2019). Consistency and variability in children's word learning across languages. *Open Mind*, **3**, 52–67. doi: https://doi.org/10.1162/opmi_a_00026

Coffey, J. R., **Shafto, C. L.**, **Geren, J. C.**, & **Snedeker, J.** (2022). The effects of maternal input on language in the absence of genetic confounds: Vocabulary development in internationally adopted children. *Child Development*, **93**(1), 237–253. doi: https://doi.org/10.1111/cdev.13688.

Coffey, J. R., & **Snedeker, J.** (2024). Disentangling the roles of age and knowledge in early language acquisition: A fine-grained analysis of the vocabularies of infant and child language learners. *Cognitive Psychology*, **153**, 101681. https://doi.org/10.1016/j.cogpsych.2024.101681

Coffey, J. R., **Zeitlin, M.**, **Crawford, J.**, & **Snedeker, J.** (2024). It's all in the interaction: Early acquired words are both frequent and highly imageable. *Open Mind*, **8**, 309–332. https://doi.org/10.1162/opmi_a_00130

Dailey, S., & **Bergelson, E.** (2022). Language input to infants of different socioeconomic statuses: A quantitative meta-analysis. *Developmental Science*, **25**(3), e13192–n/a. https://doi.org/10.1111/desc.13192

Dupas, P., **Falezan, C.**, **Jayachandran, S.**, & **Walsh, M. P.** (2023). Informing mothers about the benefits of conversing with infants: Experimental evidence from Ghana. *National Bureau of Economic Research Working Paper Series*, **31264**. http://www.nber.org/papers/w31264

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. https://doi.org/10.1136/bmj.315.7109.629

Gauthier, K., **Genesee, F.**, **Dubois, M. E.**, & **Kasparian, K.** (2013). Communication patterns between internationally adopted children and their mothers: Implications for language development. *Applied PsychoLinguistics*, **34**(2), 337–359. https://doi.org/10.1017/S0142716411000725

Gelman, A., & **Carlin, J.** (2014). Beyond power calculations: Assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, **9**(6), 641–651. https://doi.org/10.1177/1745691614551642

Gilkerson, J., **Richards, J. A.**, **Warren, S. F.**, **Montgomery, J. K.**, **Greenwood, C. R.**, **Kimbrough Oller, D.**, **Hansen, J. H. L.**, & **Paul, T. D.** (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, **26**(2), 248–265. https://doi.org/10.1044/2016_AJSLP-15-0169

Gillette, J., **Gleitman, H.**, **Gleitman, L.**, & **Lederer, A.** (1999). Human simulations of vocabulary learning. *Cognition*, **73**(2), 135–176. https://doi.org/10.1016/S0010-0277(99)00036-0

Golinkoff, R. M., **Hoff, E.**, **Rowe, M. L.**, **Tamis-LeMonda, C. S.**, & **Hirsh-Pasek, K.** (2019). Language Matters: Denying the Existence of the 30-Million-Word Gap Has Serious Consequences. *Child Development*, **90**(3), 985–992. https://doi.org/10.1111/cdev.13128

Hardy-Brown, K., **Plomin, R.**, & **DeFries, J. C.** (1981). Genetic and environmental influences on the rate of communicative development in the first year of life. *Developmental Psychology*, **17**(6), 704–717. doi: https://doi.org/10.1037/0012-1649.17.6.704

Hart, B., & **Risley, T. R.** (1995). *Meaningful differences in the everyday experience of young American children*. P.H. Brookes.

Hedges, L., & **Olkin, I.** (1985). *Statistical models for metaanalysis*. New York, NY: Academic Press.

Hirsh-Pasek, K., **Adamson, L. B.**, **Bakeman, R.**, **Owen, M. T.**, **Golinkoff, R. M.**, **Pace, A.**, **Yust, P. K.**, & **Suma, K.** (2015). The contribution of early communication quality to low-income children's language success. *Psychological Science*, **26**(7), 1071–1083. doi: https://doi.org/10.1177/0956797615581493

Hoff, E. (2003). The Specificity of Environmental Influence: Socioeconomic Status Affects Early Vocabulary Development Via Maternal Speech. *Child Development*, **74**(5), 1368–1378. https://doi.org/10.1111/1467-8624.00612

Hoff-Ginsberg, E. (1986). Function and Structure in Maternal Speech. *Developmental Psychology*, **22**(2), 155–163. https://doi.org/10.1037/0012-1649.22.2.155

Hsu, N., **Hadley, P. A.**, & **Rispoli, M.** (2017). Diversity matters: Parent input predicts toddler verb production. *Journal of Child Language*, **44**(1), 63–86. https://doi.org/10.1017/S0305000915000690

Hurtado, N., **Marchman, V. A.**, & **Fernald, A.** (2008). Does input influence uptake? Links between maternal talk, processing speed and vocabulary size in Spanish-learning children. *Developmental Science*, **11**(6), F31–F39. doi: https://doi.org/10.1111/j.1467-7687.2008.00768.x

Huttenlocher, J., **Haight, W.**, **Bryk, A.**, **Seltzer, M.**, & **Lyons, T.** (1991). Early Vocabulary Growth. *Developmental Psychology*, **27**(2), 236–248. https://doi.org/10.1037/0012-1649.27.2.236

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

Huttenlocher, J., **Vasilyeva, M.**, **Cymerman, E.**, & **Levine, S.** (2002). Language input and child syntax. *Cognitive Psychology*, **45**, 337–374. doi: https://doi.org/10.1016/S0010-0285(02)00500-5

Huttenlocher, J., **Waterfall, H.**, **Vasilyeva, M.**, **Vevea, J.**, & **Hedges, L. V.** (2010). Sources of variability in children's language growth. *Cognitive Psychology*, **61**(4), 343–365. https://doi.org/10.1016/j.cogpsych.2010.08.002

Kidd, E., & **Garcia, R.** (2022). How diverse is child language acquisition research? *First Language*, **42**(6), 703–735. https://doi.org/10.1177/01427237211066405

Leech, K. A., & **Rowe, M. L.** (2014). A comparison of preschool children's discussions with parents during picture book and chapter book reading. *First Language*, **34**(3), 205–226. https://doi.org/10.1177/0142723714534220

Loukatou, G., **Scaff, C.**, **Demuth, K.**, **Cristia, A.**, & **Havron, N.** (2022). Child-directed and overheard input from different speakers in two distinct cultures. *Journal of Child Language*, **49**(6), 1173–1192. https://doi.org/10.1017/S0305000921000623

Mannle, S., **Barton, M.**, & **Tomasello, M.** (1992). Two-year-olds' conversations with their mothers and preschool-aged siblings. *First Language*, **12**(34), 57–71. https://doi.org/10.1177/014272379201203404

Mastin, J. D., & **Vogt, P.** (2016). Infant engagement and early vocabulary development: A naturalistic observation study of Mozambican infants from 1;1 to 2;1. *Journal of Child Language*, **43**(2), 235–264. https://doi.org/10.1017/S0305000915000148

Murray, A. D., **Johnson, J.**, & **Peters, J.** (1990). Fine-tuning of utterance length to preverbal infants: Effects on later language development. *Journal of Child Language*, **17**(3), 511–525. https://doi.org/10.1017/S0305000900010862

Newport, E. L., & **Gleitman, H.** (1977). Maternal Self-Repetition and the Child's Acquisition of Language. *Papers and Reports on Child Language Development*, **13**(Aug), 46–55.

Page, M. J., **McKenzie, J. E.**, **Bossuyt, P. M.**, **Boutron, I.**, **Hoffmann, T. C.**, **Mulrow, C. D.**, **Shamseer, L.**, **Tetzlaff, J. M.**, & **Moher, D.** (2021). Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology*, **134**, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

Pan, B. A., **Rowe, M. L.**, **Spier, E.**, & **Tamis-Lemonda, C.** (2004). Measuring productive vocabulary of toddlers in low-income families: Concurrent and predictive validity of three sources of data. *Journal of Child Language*, **31**(3), 587–608. https://doi.org/10.1017/S0305000904006270

Pancsofar, N., & **Vernon-Feagans, L.** (2006). Mother and father language input to young children: Contributions to later language development. *Journal of Applied Developmental Psychology*, **27**(6), 571–587. doi: https://doi.org/10.1016/j.appdev.2006.08.003

Parker, M. D., and **Brorson, K.** (2005). A comparative study between mean length of utterance in morphemes (MLUm) and mean length of utterance in words (MLUw). *First Language* **25**, 365–376. doi: https://doi.org/10.1177/0142723705059114

Peters, J. L., **Sutton, A. J.**, **Jones, D. R.**, **Abrams, K. R.**, & **Rushton, L.** (2008). Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry. *Journal of Clinical Epidemiology*, **61**(10), 991–996. https://doi.org/10.1016/j.jclinepi.2007.11.010

Polderman, T. J., **Benyamin, B.**, **De Leeuw, C. A.**, **Sullivan, P. F.**, **Van Bochoven, A.**, **Visscher, P. M.**, & **Posthuma, D.** (2015). Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, **47**(7), 702–709. doi: https://doi.org/10.1038/ng.3285

Pustejovsky, J. (2024). *clubSandwich*: Cluster-Robust (Sandwich) Variance Estimators with Small-Sample Corrections. R package version 0.5.11. http://jepusto.github.io/clubSandwich/.

Pustejovsky, J.E. & **Tipton, E.** (2022). Meta-analysis with Robust Variance Estimation: Expanding the Range of Working Models. *Preventative Science* **23**, 425–438. https://doi.org/10.1007/s11121-021-01246-3

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Rodgers, M. A., & **Pustejovsky, J. E.** (2021). Evaluating meta-analytic methods to detect selective reporting in the presence of dependent effect sizes. *Psychological Methods*, **26**(2), 141–160. https://doi.org/10.1037/met0000300

Romeo, R. R., **Leonard, J. A.**, **Robinson, S. T.**, **West, M. R.**, **Mackey, A. P.**, **Rowe, M. L.**, & **Gabrieli, J. D. E.** (2018). Beyond the 30-Million-Word Gap: Children's Conversational Exposure Is Associated With Language-Related Brain Function. *Psychological Science*, **29**(5), 700–710. https://doi.org/10.1177/0956797617742725

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

Rosenthal, R. (1979). The 'File Drawer Problem' and tolerance for null results. *Psychological Bulletin* **86**, 638–641.

Rowe, M. L. (2008). Child-directed speech: Relation to socioeconomic status, knowledge of child development and child vocabulary skill. *Journal of Child Language*, **35**(1), 185–205. https://doi.org/10.1017/S0305000907008343

Rowe, M. L. (2012). A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. *Child Development*, **83**(5), 1762–1774. https://doi.org/10.1111/j.1467-8624.2012.01805.x

Schieffelin, B. B., & **Ochs, E.** (1986). *Language socialization across cultures*. Cambridge University Press.

Shneidman, L. A., & **Goldin-Meadow, S.** (2012). Language input and acquisition in a Mayan village: How important is directed speech? *Developmental Science*, **15**(5), 659–673. https://doi.org/10.1111/j.1467-7687.2012.01168.x

Sperry, D. E., **Sperry, L. L.**, & **Miller, P. J.** (2019). Reexamining the verbal environments of children from different socioeconomic backgrounds. *Child Development*, **90**(4), 1303–1318. https://doi.org/10.1111/cdev.13072

Stromswold, K. (2001). The heritability of language: A review and meta-analysis of twin, adoption, and linkage studies. *Language*, **77**(4), 647–723. doi: https://doi.org/10.1353/lan.2001.0247

Suskind, D. L., **Leffel, K. R.**, **Graf, E.**, **Hernandez, M. W.**, **Gunderson, E. A.**, **Sapolich, S. G.**, **Suskind, E.**, **Leininger, L.**, **Goldin-Meadow, S.**, & **Levine, S. C.** (2016). A parent-directed language intervention for children of low socioeconomic status: A randomized controlled pilot study. *Journal of Child Language*, **43** (2), 366–406. https://doi.org/10.1017/S0305000915000033

Tamis-LeMonda, C. S., **Kuchirko, Y.**, **Luo, R.**, **Escobar, K.**, & **Bornstein, M. H.** (2017). Power in methods: Language to infants in structured and naturalistic contexts. *Developmental Science*, **20**(6), e12456. https://doi.org/10.1111/desc.12456

Tsuji, S., **Cristia, A.**, **Frank, M. C.**, & **Bergmann, C.** (2020). Addressing publication bias in meta-analysis: Empirical findings from community-augmented meta-analyses of infant language development. *Zeitschrift für Psychologie*, **228**(1), 50. https://doi.org/10.1027/2151-2604/a000393

Turkheimer, E., **Haley, A.**, **Waldron, M.**, **D'Onofrio, B.**, & **Gottesman, I. I.** (2003). Socioeconomic status modifies heritability of IQ in young children. *Psychological Science*, **14**(6), 623–628. https://doi.org/10.1046/j.0956-7976.2003.psci_1475.x

Vembye, M. H., **Pustejovsky, J. E.**, & **Pigott, T. D.** (2023). Power approximations for overall average effects in meta-analysis with dependent effect sizes. *Journal of Educational and Behavioral Statistics*, **48**(1), 70–102. https://doi.org/10.3102/10769986221127379

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

Wadsworth, S. J., **Corley, R. P.**, **Hewitt, J. K.**, **Plomin, R.**, & **DeFries, J. C.** (2002). Parent–offspring resemblance for reading performance at 7, 12 and 16 years of age in the Colorado Adoption Project. *Journal of Child Psychology and Psychiatry*, **43**(6), 769–774. https://doi.org/10.1111/1469-7610.00085

Walker, D., **Greenwood, C.**, **Hart, B.**, & **Carta, J.** (1994). Prediction of school outcomes based on early language production and socioeconomic factors. *Child development*, **65**(2), 606–621. https://doi.org/10.1111/j.1467-8624.1994.tb00771.x

Wang, Y., **Williams, R.**, **Dilley, L.**, & **Houston, D. M.** (2020). A meta-analysis of the predictability of LENA™ automated measures for child language development. *Developmental Review*, **57**, 100921. https://doi.org/10.1016/j.dr.2020.100921

Weber, A., **Fernald, A.**, & **Diop, Y.** (2017). When cultural norms discourage talking to babies: Effectiveness of a parenting program in rural senegal. *Child Development*, **88**(5), 1513–1526. https://doi.org/10.1111/cdev.12882

Weisleder, A., & **Fernald, A.** (2013). Talking to children matters: Early language experience strengthens processing and builds vocabulary. *Psychological Science*, **24**(11), 2143–2152. https://doi.org/10.1177/0956797613488145

Wong, K., **Thomas, C.**, & **Boben, M.** (2020). Providence talks: A citywide partnership to address early childhood language development. *Studies in Educational Evaluation*, **64**, 100818. https://doi.org/10.1016/j.stueduc.2019.100818

https://doi.org/10.1017/S0305000924000692 Published online by Cambridge University Press

Zauche, L. H., **Thul, T. A.**, **Mahoney, A. E. D.**, & **Stapel-Wax, J. L.** (2016). Influence of language nutrition on children's language and cognitive development: An integrated review. *Early Childhood Research Quarterly*, **36**, 318–333. https://doi.org/10.1016/j.ecresq.2016.01.015

Zhang, X., **Liu, D.**, **Pappas, L.**, **Dill, S. E.**, **Feng, T.**, **Zhang, Y.**, **Zhao, J.**, **Rozelle, S.**, & **Ma, Y.** (2023). The home language environment and early childhood development: A LENA study from rural and peri-urban China. *Applied Developmental Science*, 1–19. https://doi.org/10.1080/10888691.2023.2267440

Cite this article: Coffey, J.R., & Snedeker, J. (2025). How strong is the relationship between caregiver speech and language development? A meta-analysis. *Journal of Child Language* 1–36, https://doi.org/10.1017/S0305000924000692