This is the peer reviewed version of the following article, which has been published in final form on 24 May 2025 at https://doi.org/10.1002/icd.70019. This article may be used for non-commercial purposes in accordance with Wiley Terms and Conditions for Use of Self-Archived Versions. This article may not be enhanced, enriched or otherwise transformed into a derivative work, without express permission from Wiley or by statutory rights under applicable legislation. Copyright notices must not be removed, obscured or modified. The article must be linked to Wiley's version of record on Wiley Online Library and any embedding, framing or otherwise making available the article or pages thereof by third parties from platforms, services and websites other than Wiley Online Library must be prohibited.

How well do children remember fast-mapped words? A pre-registered meta-analysis of retention following the mutual exclusivity response

Stage 2 Registered Report meta-analysis

Emily Mather

Shane Lindsay

School of Psychology and Social Work, University of Hull, UK.

Author corrected manuscript accepted at *Infant and Child Development*

Keywords: language development; word learning; fast mapping; mutual exclusivity; meta-analysis; registered report

Abstract

There is widespread evidence that children display a mutual exclusivity response upon encountering new words. Children displaying this behaviour will select a novel, name-unknown object in response to a novel label, rather than a familiar, name-known object. The mutual exclusivity response has been viewed as a means of fast-mapping vocabulary– enabling the retention of new words with minimal and incidental exposure. Thus, it may play an important role in driving early vocabulary growth. Yet, while the mutual exclusivity response may initially guide the correct choice of a novel word's referent, it does not necessarily result in retention, for even brief durations. In this pre-registered meta-analysis we examined the evidence for retention of novel noun mappings disambiguated through the mutual exclusivity response. While a large effect size was observed for the mutual exclusivity response, consistent with past research, there was a smaller effect for subsequent retention– which was substantially attenuated by Bayesian publication bias correction. Our findings are consistent with models of word learning in which referent selection is not governed by the same processes underlying referent retention.

Summary

- In this meta-analysis, we examined whether the mutual exclusivity response leads to the retention of object labels.
- We observed a large effect size for referent selection (the mutual exclusivity response), contrasting with a smaller effect size for referent retention.
- Age was a significant positive moderator of both referent selection and referent retention.
- The referent retention effect size was considerably attenuated by publication bias correction.
- Our findings are consistent with models of word learning where referent selection and referent retention are distinct processes.

1. Introduction

A key challenge for language development researchers is to understand how children build a vocabulary. Decades of research have made progress towards understanding how children solve the puzzle of correctly assigning words to referents (Quine, 1960). Models of lexical disambiguation identify a range of social (Bloom, 2000), associative (Smith & Yu, 2008; McMurray, Horst, Samuelson, 2012), and lexical (Golinkoff, Mervis, & Hirsk-Pasek, 1994) cues available to children at different points in development and which may interact (e.g., Yurovsky & Frank, 2017). However, the correct identification of a referent is only evidence of a child's ability to use a word learning cue; the response is not necessarily evidence of a new entry in the child's lexicon. Contrary to earlier arguments that children can 'fast map' vocabulary (Markson & Bloom, 1997), the literature increasingly demonstrates the fragility of children's memory for novel word mappings (Horst & Samuelson, 2008; Vlach, 2019). Here we focus on one word learning behaviour, commonly known as *mutual exclusivity*, to review and systematically analyse the evidence that children can retain word mappings formed on this basis. We do so through the use of a pre-registered meta-analysis, following calls for the greater use of open science methods in developmental research (see Bergmann et al., 2018; Davis-Kean & Ellis, 2019).

Mutual exclusivity is a means of disambiguating reference in the absence of ostensive labelling. A child displaying the mutual exclusivity response will choose a novel, name-unknown object as the referent of a novel label as opposed to a familiar, name-known object. We use the term 'mutual exclusivity response' to distinguish the behaviour from its many differing explanations. For example, the child rejects multiple labels for objects (Markman, 1989, 1990), the child seeks nameless objects as referents for novel labels (Mervis & Bertrand, 1994; Golinkoff, Mervis, & Hirsh-Pasek, 1994), the child is biased towards novelty (Mather & Plunkett, 2012; Horst, Samuelson, Kucker, & McMurray, 2011), or is engaged in social-pragmatic reasoning (Grassman, Stracke, & Tomasello, 2009). Many studies have demonstrated children's use of mutual exclusivity in word learning (e.g., Golinkoff, Hirsh-Pasek, Bailey, & Wegner, 1992; Wilkinson, Ross, & Diamond, 2003) with the onset of the response occurring around 15 to 17 months of age (Halberda, 2003; Markman, Wasow, & Hansen, 2003; though see Bion, Borovsky, & Fernald, 2013), with evidence of behaviour related to mutual exclusivity as early as 10 months of age (Dewar & Zu, 2007; Mather & Plunkett, 2010).

The mutual exclusivity response was initially viewed as a means of 'fast mapping' vocabulary (Carey & Bartlett, 1978) – that new words can be learnt through minimal, incidental exposure and retained long-term. If mutual exclusivity accelerates word learning then it may drive the purported vocabulary spurt during the second year of life (e.g., Markman, Wasow, & Hansen, 2003). An alternative view is that the mutual exclusivity response is an outcome of this period of vocabulary growth (see Mervis & Bertrand, 1994). A recent meta-analysis of the mutual exclusivity response by Lewis et al. (2020) found a large overall effect size, with both age and productive vocabulary as significant positive (though highly intercorrelated) predictors. Lewis et al. further report two mutual exclusivity experiments using a touchscreen object selection task with two- to four-year-olds. In the first experiment, vocabulary size predicted accuracy of the mutual exclusivity response with the addition of age producing no significant predictive improvement. In the second experiment, systematically manipulating exposure to the familiar object-label mapping (through prior training) resulted in improvements in the mutual exclusivity response. These combined findings suggest that the robustness of the response develops across the preschool years, driven by linguistic experience.

In principle, the mutual exclusivity response could enable word mapping through a single exposure and in the absence of ostensive naming. However, only a subset of studies investigating mutual exclusivity test for the subsequent retention of novel word mappings. These studies include *referent selection* trials to test for the mutual exclusivity response and *retention* trials to probe memory for the novel word mappings. During retention trials, novel referents are presented together and labelled. Unlike children's performance on referent selection trials, successful responding at test cannot rely on the mutual exclusivity response. Instead, correct responses require knowledge of the specific word-referent mappings (see Schafer & Plunkett, 1998). Through the inclusion of retention trials, there is now conflicting evidence that mutual exclusivity supports retention of novel word mappings. An object selection study by Horst and Samuelson (2008) found that while two-year-olds were able to apply mutual exclusivity during a referent selection task, they were unable to retain the novel word-object mappings across a five minute interval. The children required supplementary ostensive naming to succeed in the retention task. The mutual exclusivity response may also developmentally lag behind unambiguous labelling in enabling vocabulary retention. In a preferential looking task, Bion, Fernald, and Borovsky (2013) found that while 18-month-olds were able to retain novel-word object mappings labelled ostensively (i.e., in the absence

of distractor objects), only by 30 months of age were children able to retain the novel mappings when disambiguated through mutual exclusivity. While other studies do provide evidence of retention via mutual exclusivity, these studies also suggest retention is sensitive to properties of the referent selection task, such as familiar distractor salience (Pomper & Saffran, 2019), lexical strength (Kucker, McMurray, & Samuelson, 2020) and repetition (Axelsson & Horst, 2014), and the variability of novel object exemplars (Twomey, Ranson, & Horst, 2014). These findings call into question the role of mutual exclusivity in vocabulary development.

How can children reliably select the correct referent of a new word, only to forget the mapping minutes later? And what, therefore, underlies their vocabulary development? If children are successful at initially identifying the referents of words, then the bottleneck may lie with the encoding and retention of word mappings. One recent view is that referent selection and associative memory are dissociable processes operating on different timescales. Under McMurray, Horst, and Samuelson's (2012) associative model of word learning, referent selection is driven by competitive activation processes distinct from slower associative learning processes occurring over developmental time. Importantly, the response of selecting the correct referent may only lead to weak learning of the association. Thus, further exposures to the word-object association are required. McMurray et al. argue that mutual exclusivity has previously been conceptualised as a constraint on *learning*, for example, that it is difficult to learn second labels for objects. However, in the face of evidence that mutual exclusivity does not necessarily lead to retention of novel word mappings, it is perhaps better understood as an 'in the moment' response during communication (see also Kucker, McMurray, & Samuelson, 2015).

In sum, given the discrepancies found between children's referent selection and retention abilities, it is crucial that we more precisely understand the effect of mutual exclusivity on retention and how this may systematically vary with properties of the task and participants. In turn, this will help us better understand the nature of early vocabulary development and constrain models of word learning. The meta-analysis by Lewis et al. (2020) examined the mutual exclusivity response but not subsequent retention. Therefore, we will undertake a meta-analysis of word learning literature to measure the strength of retention for word mappings disambiguated through mutual exclusivity. We focus on children aged one to five years, using evidence from two commonly used methods in the literature: object

selection and intermodal preferential looking tasks. In object selection studies, referent selection and retention is measured through the child's selection (through pointing or touching) of an object as the referent of a label. In intermodal preferential looking (IPL) studies, referent selection and retention is measured through a child's preference to visually fixate an object in response to a label.

This meta-analysis is pre-registered as a Stage 1 registered report following best practices of open science. Given the uncertainty of the evidence for retention following the display of the mutual exclusivity response, our primary goal is to evaluate across studies the strength of evidence for retention of word mappings following the mutual exclusivity response. Along with this primary goal, we consider a non-exhaustive set of potential moderators of the retention effect. These variables were chosen for theoretical interest and where variability across studies appears likely. However, as a pre-registered meta-analysis, we cannot ascertain in advance how many studies will be available for moderator inclusion. Consequently, as with any meta-analysis, statistical power to test these moderators will vary and hence informativeness of any conclusions. Nevertheless, we consider this a good candidate set of variables. Even where evidence is limited, the analysis will highlight where further research is needed into the effect of these variables on retention following the mutual exclusivity response.

1.1 Potential Moderators of the Retention Effect Size

1.1.1 Age. There are age related improvements in the robustness of the mutual exclusivity response (Lewis et al., 2020) and retention of novel word mappings (Bion et al., 2013). Therefore, we expect age to predict the retention of mappings formed through the mutual exclusivity response.

1.1.2 Vocabulary. Like age, vocabulary is associated with improvements in the mutual exclusivity response and subsequent retention (see Bion et al., 2013; Lewis et al., 2020). Therefore, we also expect vocabulary to positively predict retention. However, the inclusion of a vocabulary moderator may be constrained by limited availability of data (Lewis et al., 2020, were only able to include 18% of total studies in their vocabulary moderator analysis). We also note that the high intercorrelation of age and vocabulary measures (see Lewis et al., 2020) means they need to be examined together.

1.1.3 Number of Novel Label Presentations. If the mutual exclusivity response is not an outcome of rapid word mapping, then increased opportunities to associate the novel label with the novel object may facilitate retention (see McMurray et al., 2012). We expect that the number of presentations of each novel label mapping (in the context of a mutual exclusivity trial) will be a positive predictor of retention.

1.1.4 Mutual Exclusivity Effect Size. We expect to observe a similarly large mutual exclusivity effect size to Lewis et al. (2020), given that our sample will mostly be a subset of those included in their analysis, although our smaller sample will reduce statistical power. As a moderator, greater attention to the target referent during mutual exclusivity trials could provide greater opportunity to form a mapping that is retained (see Bion et al., 2013). However, the presence of studies demonstrating a mutual exclusivity response without subsequent retention (e.g., Horst and Samuelson, 2008) means it is unclear whether the size of the mutual exclusivity effect will be predictive of retention.

1.1.5 Response Measure. The retention effect may be moderated by the response measure - preferential looking or object selection. Gurteen et al. (2011) found evidence for retention of ostensively labelled novel objects in 13- and 17-month-olds when measured by target looking but not when measured by reaching to target. This finding suggests that the mutual exclusivity retention effect may be stronger for preferential looking tasks than object selection tasks.

1.1.6 Multilingualism. While the mutual exclusivity response appears weaker in multilingual children (e.g., Byers-Heinlein & Werker, 2009; Houston-Price et al., 2010) there is currently mixed evidence for weaker retention following the mutual exclusivity response in multilingual children (Kalashnikova, Escudero, & Kidd, 2018; Weatherhead, Arredondo, Nácar Garcia, & Werker, 2021). Therefore, it is unclear whether the number of languages spoken will moderate the retention effect size.

1.1.7 Number of Novel Mappings. The number of novel mappings to learn varies across mutual exclusivity studies of retention, from two mappings in Bion et al (2013) to eight mappings in Gambi et al (2021). We expect that the number of novel mappings to learn will impact the memory load of the word learning task, and therefore expect a negative relationship with retention.

1.1.8 Referent Selection Distractors. The number of familiar distractors per referent selection trial varies across retention studies of mutual exclusivity. For example, referent selection trials might pair a novel target with either one (Bion et al., 2013; Pomper & Saffran, 2019) or two (Twomey et al., 2014; Kucker et al., 2018) familiar distractors per trial. Horst et al. (2010) found that increasing the number of familiar distractors in a mutual exclusivity task was detrimental to retention. Although Horst et al. provided supplementary word-learning cues beyond mutual exclusivity, their findings suggest that additional familiar distractors will be negatively related to retention effect size.

1.1.9 Referent Retention Distractors. The number of novel objects presented during individual retention trials may vary between methodologies. Intermodal preferential looking studies typically present two previously labelled novel objects during retention trials, one target and one distractor (see Schafer & Plunkett, 1998). Preferential looking studies of mutual exclusivity similarly present one novel distractor per retention trial (e.g., Kalashnikova et al., 2018; Pomper & Saffran, 2019). However, object selection studies often present more than one distractor within retention trials (e.g., Twomey et al., 2014; Kucker et al., 2018). We expect increased processing demands with additional distractors and therefore a negative predictive relationship with retention.

1.1.10 Retention Interval. Forgetting of newly learnt mappings is more likely with a delay between referent selection and referent retention trials (see Vlach, 2019). Mutual exclusivity studies typically test retention either immediately after the completion of referent selection trials (e.g., Kalashnikova et al., 2018; Pomper & Saffran, 2019) or following a five-minute interval (e.g., Horst & Samuelson, 2008; Gambi et al., 2021). We expect to observe a negative predictive relationship between the retention duration and the retention effect size.

2. Method

2.1 Eligibility Criteria

Our focus is on noun learning studies using the IPL or object selection paradigm. For a study to be included, the primary criterion was to have a condition in the study with the use of mutual exclusivity in referent selection trials followed by a retention test. These referent selection trials presented at least one familiar, name-known object and only one novel, nameless object as possible referents of a novel object label. Only the use of mutual exclusivity allowed disambiguation of the novel label. This required exclusion of studies with

other means of identifying the novel referent, such as cues identifying the target (e.g., eye gaze or pointing, movement or perceptual salience of target), statistical information (i.e., only the target co-occurs with the novel label across trials) or a physical transformation which excludes distractors (e.g., removal or reduced visibility of distractors). Studies were also excluded if the child's response was followed by ostensive labelling or highlighting of the target, corrective feedback, or positive reinforcement of target selections. Retention trials required at least two previously labelled novel objects presented together and labelled. The novel object exemplars at test were identical to those presented during referent selection trials.

We will focus on studies with children aged 1 – 5 years. We expect most studies will test children under three years of age, but we wanted to include studies from pre-school children (whose age threshold varies across countries) therefore used an upper age range of 5 years. Only typically developing children will be included to prevent influences of sensory, cognitive, or physical impairment on word mapping and retention. This means exclusion of children with visual or hearing impairments and children from atypically developing populations (e.g., pre-term, language delay, autism spectrum disorder, specific language impairment). We have also excluded late talkers if specifically included as a comparison group. While we expect the majority of studies to use monolingual populations, we will include studies with multilingual participants if they occur.

2.2 Selection Procedure

We restricted our search to studies published in English. Our initial search used EBSCO with the following databases: Academic Search Premier; CINAHL Complete; Education Research Complete; MEDLINE; APA PsycArticles; APA PsycInfo. The search criteria were: TI((Word learning OR fast mapping OR Noun learning OR Mutual Exclusivity OR Referent selection) AND (Retention OR Memory) NOT (verb OR adjective OR disorder OR systematic review) OR AB(Word learning OR fast mapping OR Noun learning OR Mutual Exclusivity OR Referent selection) AND (Retention OR Memory) NOT (verb OR adjective OR disorder OR systematic review). This search yielded a total of 2161 records for screening.

To further populate our primary search results, we ran citation searches for the two most cited papers known to us on mutual exclusivity and retention: Horst & Samuelson (2008) and Bion et al. (2013). We used the Cited By search report from Web of Science as

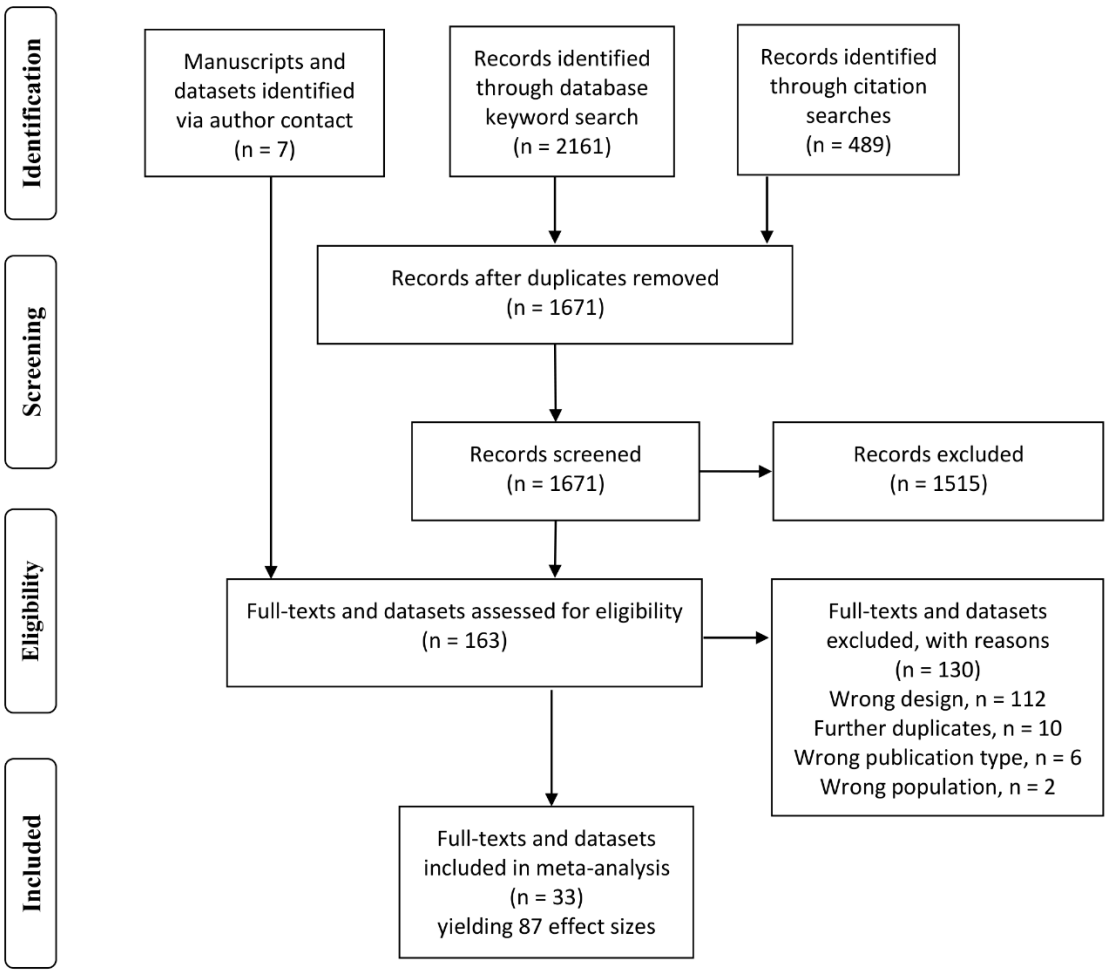
our main citation search. This was supplemented with the Cited By search results from Google Scholar for greater coverage of grey literature including conference papers, theses, and preprints compared to EBSCO and Web of Sciences searches (Walters, 2007). These Web of Science and supplementary Google Scholar citation searches yielded a combined total of 489 records.

2.3 Screening and Coding

Following removal of duplicates across database and citation searches, there were a total of 1671 records for abstract screening. Screening was facilitated through the Rayyan tool (Ouzzani et al., 2016), with decisions made within Rayyan imported into an inclusion worksheet accessible on the project OSF repository. Two members of the research team independently screened all abstracts for initial inclusion, leading to the selection of 156 records. Following abstract screening, each full text article was surveyed by a research team member to comprehensively determine eligibility based on our inclusion criteria. Sources deemed not suitable are documented with a code to indicate the reason for exclusion. To identify any other eligible unpublished or pre-publication data, we issued a call on the Cognitive Development Society (CDS) listserv. We also contacted several authors who have published mutual exclusivity retention data to ask for any unpublished data that may fit the selection criteria (see OSF templates folder). This yielded a total of seven manuscripts and datasets of which five were eligible. This led to a total of 87 studies/conditions eligible for coding across 33 full-texts and datasets. A PRISMA diagram of the screening process is shown in Figure 1.

Figure 1

Meta analysis flow diagram (adapted from PRISMA; Moher et al., 2009)



Where studies met the inclusion criteria but the data were not available to calculate effect size (e.g., only reporting the difference score between pre- and post-naming), or where further clarifications were needed, we emailed authors using the template on OSF (see OSF templates folder). If we did not receive a response to our first email, we followed up two weeks later, and allowed four weeks for a response from the first email.

A coding sheet is available on the OSF repository (OSF coding folder) which describes the variables coded from each study, along with a coding key which provides an explanation and examples of how data was coded. Studies were coded by two members of the research team. Any disagreements were documented on the coding sheet and resolved by a senior research team member through verbal discussion until a consensus was reached.

For each experiment within a paper, every condition (or combination of conditions for factorial designs) was coded separately, for example, different age groups, trial types or prompts, and other manipulations. Coding includes bibliographic information about the paper and the following moderator variables: mean age of child participants (in months), mean production vocabulary measured with the Words and Sentences form of the MacArthur-Bates Communicative Development Inventory (MCDI; Fenson et al., 1994, Fenson et al., 2007) or a language adapted version, and participant population (monolingual, multilingual). We also coded a variety of methodological variables that may influence the retention effect size and observed heterogeneity, including task type (IPL or object selection), number of novel mappings to learn, total number of novel label referent selection trials, number of referent selection trials per novel mapping, number of novel label presentations per referent selection trial, number of familiar objects per referent selection trial, duration of retention interval prior to retention test, number of retention trials, and number of novel objects per retention trial. For IPL studies we coded the duration of the pre- and post-naming windows for both referent selection and referent retention trials. Our focus on novel label trials means that responses to trials presenting familiar labels or no labels were not coded.

2.4 Dependent variable

For both the referent selection and retention task, we required a measure that is meaningfully comparable across IPL and referent selection paradigms, but is also calculable from information typically reported in such studies. Initial investigation of the literature showed that there were a range of measures in use. Some studies use comparison with a chance level of performance, whereas some studies using the IPL paradigm use the difference in looking time before naming to looking times after naming. Because object selection procedures typically only measure a response following naming, our choice of dependent measure was a post-naming comparison to chance. For example, if the target object was presented alongside two distractor objects, we compared post-naming performance with a .33 chance criterion. For IPL studies, the dependent measure was the mean proportion of looking time compared with chance, whereas for object selection studies the dependent measure was the proportion of target object selections compared to chance. This choice of dependent measure means that the evidence for the mutual exclusivity effect and retention of mappings reported here may differ to that reported in an individual paper and not necessarily be consistent with the papers' conclusions (e.g., significant evidence for retention). However, this may have the

advantage of reducing the influence of publication bias. While this dependent measure could have led to the exclusion of IPL studies, a post-naming measure was codable for all but one study otherwise meeting the inclusion criteria[1].

2.5 Calculation of effect size

Effect sizes are calculated based on a one sample t-test, with the null value the chance level of performance in the task (e.g., 50% in choosing between two objects, 33% for three objects).

Effect sizes were calculated from descriptive statistics in the text, using the reported means and standard deviations, and converting from standard errors or confidence intervals where necessary. This approach was applied even when effect sizes were directly provided**.** When these statistics were unavailable, and where a t-test was reported comparing with chance performance, we converted to a Cohen's dz from the t value and df. In cases where required data was not available in the text but was present in figures, we used the WebPlotDigitizer (https://automeris.io/WebPlotDigitizer/) for calculation of means and standard deviations from error bars showing interval estimates. In several cases effect sizes were calculated from raw data available publicly or shared by the authors.

Effect sizes were initially calculated as Cohen's d using *metafor* package in R (Viechtbauer, 2010) and then converted to Hedges g for final reporting to allow an less biased estimate taking into consideration sample size.Variance of d follows the calculation used in *metafor* ((1/n) + (d^2 / (2n)). To assess the potential influence of outlier effect sizes, we conducted sensitivity analyses on the effects with winsorising of values to three standard deviations above or below that of the mean effect size for any values reaching that threshold.

2.6 Publication Bias Methodology

To provide a comprehensive analysis of publication bias, we combine qualitative and quantitative approaches. We use contour-enhanced funnel plots to visually assess the distribution of effects in the model referent selection and referent retention tasks, complementing this with a robust Bayesian model-averaging (RoBMA) approach for a quantitative evaluation of publication bias (Maier et al., 2023). Using Bayesian model-

[1] For one IPL study we were unable to obtain post-naming data and therefore the effect size calculation was based on proportional difference scores between the pre- and post-naming phases.

averaging provided a Bayes factor, quantifying evidence for or against publication bias across multiple models, rather than relying on inference from a single model (see also Hinne et al., 2020). Because no single publication bias adjustment method performs satisfactorily across all conditions, reporting and comparing multiple methods is recommended (Carter et al., 2019). RoBMA allows for a principled combination (via Bayesian model-averaging) of "naive" models, which assume no publication bias, and models that account for publication bias.

RoBMA incorporates both selection and regression models to address publication bias (Bartoš et al., 2023). Selection models adjust meta-analytic estimates based on the relative probability of studies with p-values within pre-specified intervals being published, assuming that smaller p-values are more likely to be published. We employed a two-step model assuming all significant findings ($p < .05$) are published and nonsignificant findings are published 50% of the time. We also used a three-step model assuming all significant findings are published, marginally significant findings ($p > .05$ & $p < .10$) are published 66% of the time, and nonsignificant findings 33% of the time (Bartoš et al., 2022). Regression models adjust the mean effect based on the relationship between effect sizes and standard errors, incorporating both PET (Precision Effect Test) and PEESE (Precision Effect Estimate with Standard Error; Stanley & Doucouliagos, 2014). These methods operate on the assumption that, in the absence of publication bias, there should be no relationship between effect sizes and standard errors. However, with publication bias, studies with higher standard errors are more likely to be published, leading to inflated effect size estimates.

We used the default priors in the RoBMA package, with a default effect prior of a normal distribution with a mean of 0 and SD of 1. This was truncated at 0 as for both the referent selection and retention tasks we did not have grounds to expect true negative effects. For the heterogeneity parameter we used a modestly informed prior distribution approximated by an inverse-gamma (1, 0.15), which is based on variability seen across a larger number of psychological meta-analyses (van Erp, Verhagen, Grasman, & Wagenmakers, 2017). For the selection models the prior was a cumulative sum of the Dirichlet distribution, which ensured a monotonic reduction in the probability of production as the p-value selection threshold increased. PET and PEESE used half-Cauchy distributions, which ensured a positive correlation between effect sizes and standard errors. We only

included models with random effects given our strong expectation of heterogeneity and sampling across different populations.

Our pre-registered plan specified an ensemble of eight models that allowed adjustment for publication bias: PET and PEESE comprise two models, and selection models comprise six weighting functions. Each of these eight models forms a pair - a model that assumes no effect and a version assuming an effect (with its associated prior), leading to 16 models. However, during implementation, we deviated slightly from this plan. This deviation from the pre-registered plan was made to align better with standard RoBMA practice, as the inclusion of the no-bias-adjustment models serves as a baseline for comparison and is important for properly computing the BFPB. While this represents a deviation from our pre-registered plan, we believe it enhances the robustness of our analysis without fundamentally altering its nature or introducing bias. This resulted in a total of 20 models submitted for model weighting and averaging in RoBMA. Following typical conventions, we interpreted BFPB > 3 as evidence for publication bias, BFPB < ⅓ as evidence against publication bias, and values in between as inconclusive evidence.

At the time of the pre-registration, this approach had not been developed for multilevel models and therefore does not model within-study dependencies. This limitation is true for most other quantitative methods of bias detection, and we believe despite this limitation it was the best method currently available (for comparison of performance with other methods, see Bartoš et al., 2023). This was particularly relevant in a case such as this, where heterogeneity was expected to be high and sample sizes could be small (Maier et al., 2023). Recently, however, the RoBMA approach has been adapted to account for more complicated random effects structures, including hierarchical models (Bartoš, 2025). Given the issues with our originally planned random effects structure in the frequentist model (described below), we chose to keep our original plan for a non-nested model.

3. Results

Analyses were conducted using the metafor package in R (Viechtbauer, 2010). We expected relatively high heterogeneity, given the inclusion of two experimental procedures (IPL and object selection) and various manipulations addressing different research questions. Therefore, random effects models were used. Given the inclusion of individual effect sizes where independence cannot be guaranteed, we used a multilevel model, including separate

levels for papers, and participants nested within papers (following Lewis et al., 2020). Meta-regression was conducted to examine the moderators described in the introduction. We set a minimum threshold of four studies (Schwarzer et al., 2015), each contributing at least one data point, for a moderator to be included in the final analysis. Our pre-registration stated that this would be conditional on significant values of the I^2 statistic. Upon reflection, we realised that using significant values of the Q statistic would be more appropriate, as there is no formal statistical test based on I^2 values. This decision was further reinforced by issues in the model specification that complicated the interpretation of I^2. Consequently, we modified our approach in the final analysis to rely on the significance of the Q statistic instead as our primary test of presence of heterogeneity.

3.1 Overall Effects

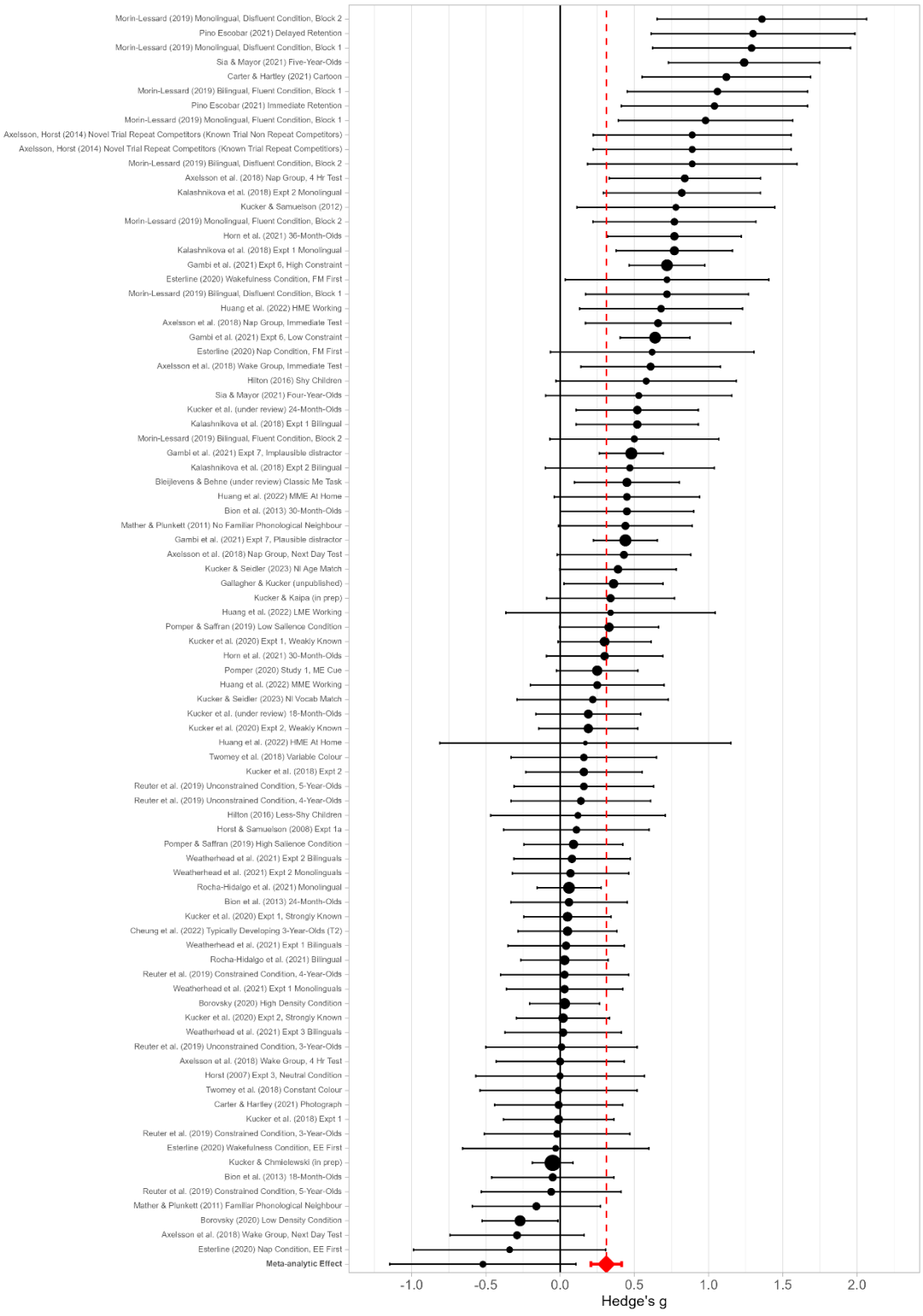
3.1.1 Referent Retention

Our main referent retention analysis (N = 87), using the pre-registered nested model to account for multiple effect sizes from the same paper, yielded an overall meta-analytic effect of Hedges' g = 0.31 [0.21, 0.41], SE = 0.053, Z = 5.92, $p < .001$. The Q test suggested significant heterogeneity ($Q(86) = 251.58$, $p < .001$), and the I^2 statistic from our pre-registered nested model was 84%.

The forest plot (Figure 2) visualises the distribution of effect sizes across studies, illustrating the variation of effects observed in our analysis (see also Tables A1 and A2).

Figure 2

Forest plot for referent retention



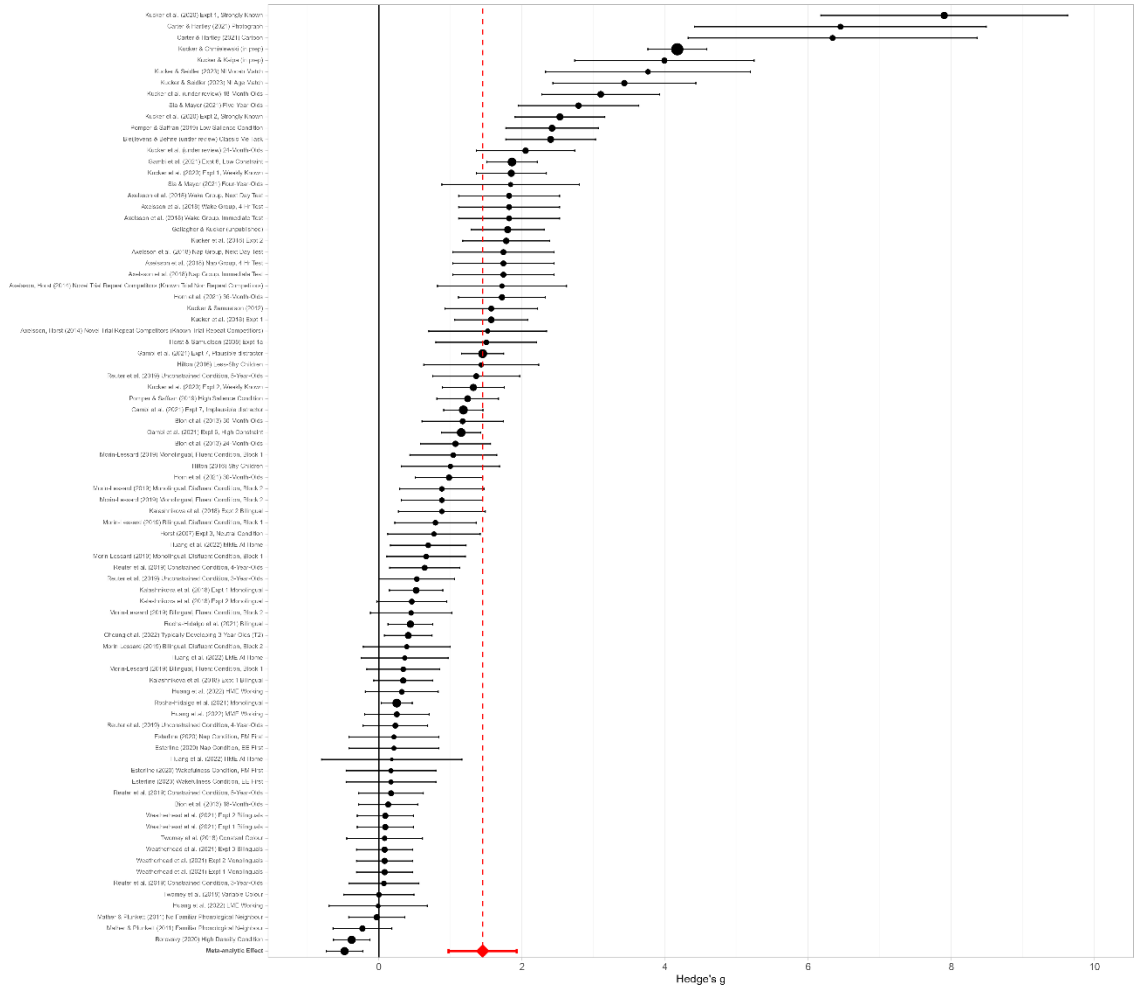
Note. Dashed line shows overall meta-analytic estimate. Error bars show standard errors.

3.1.2 Referent Selection

While our primary focus is on referent retention trials, we also analysed referent selection trials to facilitate comparison with Lewis et al. (2020). In their typically developing sample, they included 44 papers contributing 101 effect sizes. Our datasets were largely independent, with only two overlapping papers between our study and theirs. We ended up with 84 effect sizes, due to the exclusion of studies that did not include retention measures. Additionally, the final number of effects included was slightly lower than the referent retention data because three studies allowed for the calculation of a referent retention effect but not a referent selection effect.

Figure 3

Forest plot for referent selection



Note. Dashed red line overall shows meta-analytic estimate. Error bars show standard errors.

We found an overall meta-analytic effect of g = 1.45 [0.97, 1.93], p < .001 (see Figure 3). There was considerable and significant heterogeneity of effect sizes, Q = 1221.13, df = 83, p < .001, I^2 = 97%. Our findings are comparable to the overall effect size of Lewis et al. for typically-developing children d = 1.37 [1, 1.75], p < .001 for standard mutual exclusivity trials. Similarly, they also observed considerable heterogeneity (Q = 823.09, p < .001).

3.2 Moderator Analyses

3.2.1 Language Population (Monolingual vs. Multilingual)

We first examined the referent retention effect sizes for the monolingual (N = 76) and multilingual (N = 11) samples. The effect size for monolingual children, g = 0.31 [0.2, 0.42], p < .001, was similar to the effect size for multilingual children, g = 0.35 [0.14, 0.69], p = .04. A moderator analysis examining the effect of language population (monolingual vs. multilingual) on referent retention effect size was not significant, β < 0.01, Z = 0.02, p = .98. We then examined the referent selection effect size for the monolingual (N = 73) and multilingual (N = 11) samples. The effect size for monolingual children, g = 1.39 [0.91, 1.87], p < .001, was higher than the non-significant effect size for multilingual children, g = 0.97 [-0.25, 2.19], p = .11. However, a moderator analysis examining the effect of language population on the referent selection effect size was not significant (β = 0.08, Z = 0.35, p = .72). Given the lack of a significant effect of language population on either the referent retention or referent selection effect sizes, we combined the monolingual and multilingual populations for further analyses.

3.2.2 Age and Vocabulary

As expected, the most common vocabulary measure across studies was the productive vocabulary score on the Words and Sentences form of the MCDI (or a language adapted version). However, this measure was available for only a subset of the coded effects, in comparison to age as a moderator. We expected both age (N = 87) and vocabulary (N = 35) to be positive predictors of referent retention. Because age and vocabulary are highly intercorrelated (r = 0.95, df = 33, p < .001), we examined their effects in both separate and combined analyses. When analysed separately, age was a significant moderator of referent retention (β = 0.01, Z = 2.97, p = .002), but vocabulary was not (β = 0.0003, Z = 0.71, p = .48). In the combined analysis, neither age (β = -0.0384, Z = -0.68, p = .50) nor vocabulary (β = 0.0014, Z = 0.86, p = .39) were significant moderators.

For comparison with Lewis et al. (2020), we ran the same age (N = 84) and vocabulary (N = 35) moderator analyses for the referent selection effect sizes. When analysed separately, age was a significant moderator of referent selection (β = 0.03, Z = 1.98, p = 0.047), but vocabulary was not (β < 0.01, Z = 0.23, p = .82). In the combined analysis, neither age (β = 0.1164, Z = .86, p = 0.39) nor vocabulary (β = -0.0026, Z = -0.68, p = .50) were significant moderators. Our finding that age is a reliable predictor of referent selection is comparable with Lewis et al. (2020), albeit with a smaller effect. In contrast to our results, Lewis et al. (2020) found that vocabulary was still a reliable predictor when analysed separately from age.

3.2.3 Other Planned Moderators

A key question is whether or not the size of the referent selection effect is predictive of referent retention. As discussed above, more consistent attention to or selection of the target during mutual exclusivity trials may lead to better retention. Conversely, there is evidence of the mutual exclusivity response in the absence of retention (e.g., Horst & Samuelson, 2008), results captured by models of early word learning where referent selection and retention are dissociable processes (McMurray et al., 2012). We examined whether referent selection performance predicted referent retention. The analysis revealed no reliable relationship between referent selection and retention effect sizes (β = 0.01, Z = 0.44, p = 0.66), supporting the hypothesis that referent selection and retention are dissociable processes. This finding suggests that a child's ability to select novel referents in-the-moment does not necessarily predict their ability to retain those word-referent mappings over time. The absence of a reliable effect for this moderator (and others) is unlikely to be due to the prevalence of one response measure over another. Effect sizes were distributed evenly between IPL (N = 43) and object selection (N = 44) procedures (see Tables A1 and A2), and the response measure was not a significant moderator of referent retention (β = 0.09, Z = 0.78, p = .43).

Additional planned moderator analyses were also not significant. Repetition of novel labels during referent selection was expected to support retention while a greater number of novel mappings to learn was expected to impede retention. However, neither the number of novel label presentations (β = 0.03, Z = 1.3, p = .193), nor the number of novel mappings were reliable moderators (β = -0.04, Z = -1.522, p = .13), although the effects were in the predicted direction. Similarly, neither the number of distractors during referent selection trials (β = -0.13, Z = -1.24, p = .21), nor during referent retention trials (β = 0.06, Z = .73, p = .46),

were significant negative predictors of retention. Finally, the duration of the retention interval was not negatively predictive of retention (β = -0.0002, Z = -1.79, p = .074). Here use a log transform due to the large range of values with a small constant of .01 added due to the presence of zero values. The effect sizes were mainly derived from studies testing retention either immediately (N = 48) or after a five minute delay (N = 30). There were only a few effect sizes from studies with longer retention intervals of 30 minutes (N = 1), four hours (N = 2), and one day (N = 6). Table 1 provides descriptive statistics for key variables.

Table 1

Descriptive Statistics for Key Variables in Referent Selection and Retention Studies

Variable	Mean	Median	SD	Min	Max
Sample size (RR task)	27.7	20.0	26.7	4.0	209.0
Sample size (RS task)	28.3	20.0	27.5	4.0	220.0
Age (months)	31.2	29.3	10.8	16.3	66.1
Productive vocabulary size (MCDI-WS)	261.7	270.7	139.0	28.0	472.0
Number of novel mappings	3.4	3.0	1.9	1.0	8.0
Number of retention trials per novel mapping	2.6	2.0	1.8	1.0	6.0
Number of familiar distractors in selection task	1.4	1.0	0.5	1.0	2.0
Number of selection trials per familiar distractor	1.4	1.0	1.4	0.0	6.0
Total number of objects in retention task	1.6	1.0	0.7	1.0	3.0
Retention interval (minutes)	106.9	0.0	366.7	0.0	1440.0

3.2.4 Exploratory Moderators

We explored a few moderators not listed in the pre-registration. Our measure of novel label presentations (see previous section) is a combination of both the number of referent selection

trials for a given label *and* the number of label presentations within the trial. We examined whether simply the number of trials per novel label predicted retention, but found no significant effect ($\beta = 0.04$, $Z = 1.37$, $p = .17$). We also examined the effect of labelling the familiar distractors (i.e., distractors during novel label trials) during other referent selection trials. We expected that naming of the distractor objects would support the mutual exclusivity response and formation of mappings. However, neither the number of familiar distractor label presentations ($\beta = 0.02$, $Z = 1.04$, $p = .3$), nor the number of trials per familiar distractor label ($\beta = 0.03$, $Z = 0.82$, $p = .41$), was a significant moderator of retention.

3.3 Outlier Sensitivity

For referent retention, no outliers were detected using a ±3 SD criterion, while outliers were found in the selection task. To assess their impact, we conducted a sensitivity analysis by winsorising effect sizes to ±3 SD, adjusting three extreme values ($M = 6.9$, max = 8.3), likely inflated by ceiling effects and limited variability. The winsorised model yielded an effect size of 1.42 ($SE = 0.24$, 95% CI [0.96, 1.88], $z = 6.03$, $p < .001$), closely matching the original estimate of 1.45. Heterogeneity slightly decreased ($I^2 = 78.23\% \rightarrow 74.59\%$), though it remained significant ($Q(83) = 389.83 \rightarrow 347.56$, both ps < .001). These results suggest the extreme values had some influence but did not alter the significance of the effect. For further analyses described below, we included these data values in our models, and chose not to remove them on theoretical grounds.

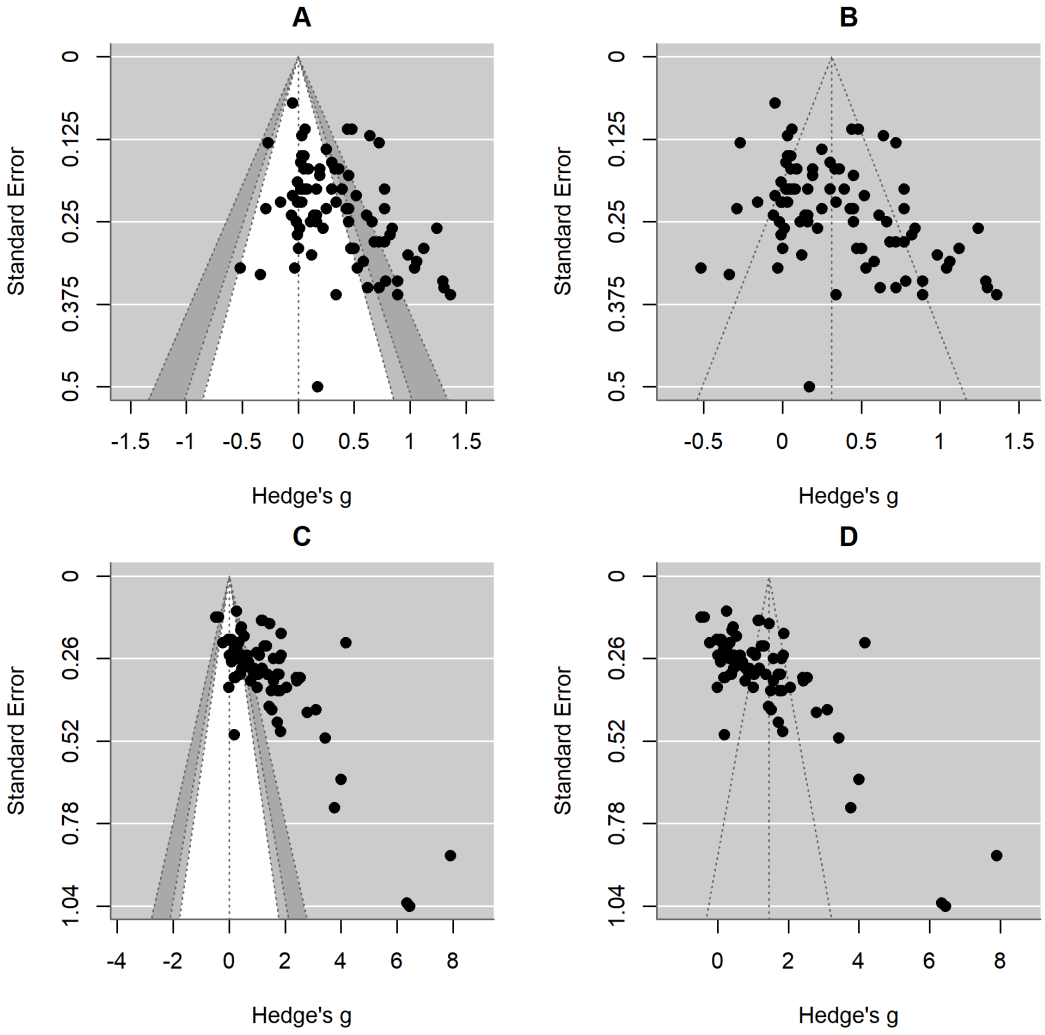
3.4 Publication Bias

3.4.1 Funnel Plots and Egger's Test

To qualitatively address publication bias and visualise the distribution of effects, we provide funnel plots in Figure 4. For both referent retention (Top, Panels A-B) and selection (Bottom, Panels C-D), we present contour-enhanced funnel plots centred on a null effect size (Panels A, C) alongside simple funnel plots with 90% confidence intervals centred on the observed meta-analytic effects (Panels B, D). The retention effects appear more symmetrical when centred on the meta-analytic effect, while the selection data show substantial skew, potentially indicating publication bias. The plots also reveal small-study effects, particularly in the selection data, where smaller studies (higher standard errors) tend to report larger effects.

Figure 4

Funnel plots of referent retention and selection.



Note. Panels A and B (top) show funnel plots for referent retention, while panels C and D (bottom) show funnel plots for Referent Selection. Panels A and C show contour-enhanced funnel plots with shading representing confidence interval regions (white: $p > .10$, grey: $.05 < p < .10$, dark grey: $p < .05$) around a null effect ($g = 0$). Panels B and D show unshaded funnel plots with a single 90% confidence interval around the observed meta-analytic effect size.

As it is common to use Egger's test to test asymmetry in funnel plots, we carried out additional exploratory analysis with this method. For referent retention, the test yielded a significant positive slope ($\beta = 2.05$, $SE = 0.7433$, $z = 2.76$, $p < .001$), indicating a positive

relationship between effect sizes and their standard errors. For referent selection, the test also revealed a significant positive slope, but with a larger magnitude (β = 6.90, *SE* = 0.785, z = 8.8, $p < .001$).

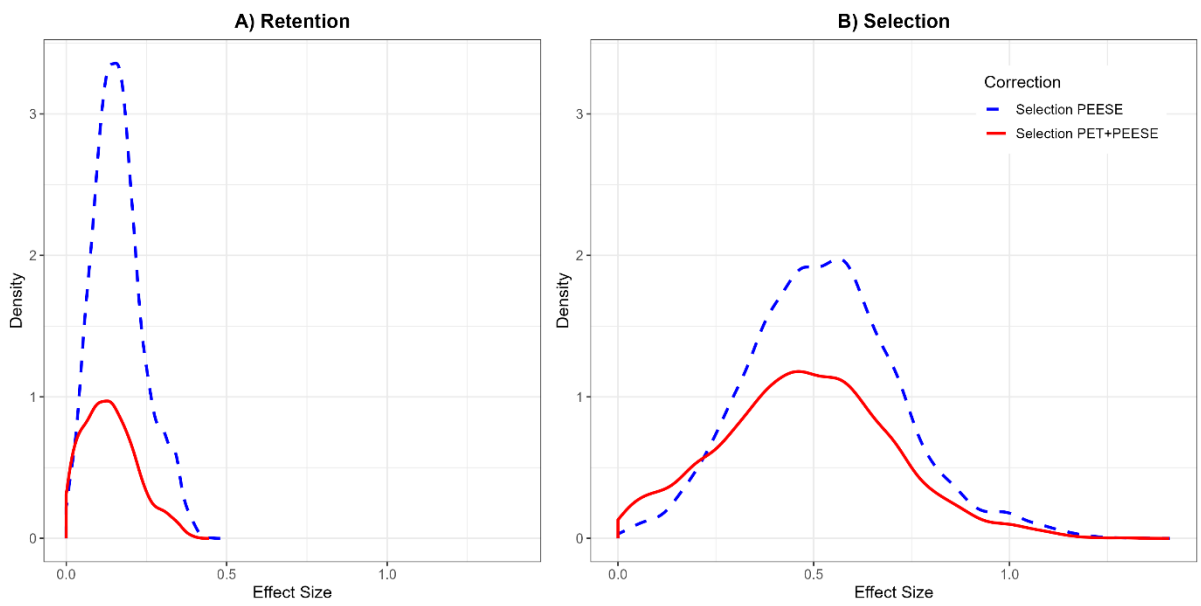
3.4.2 Bayesian Publication Bias Analysis

3.4.2.1 Referent Retention

A robust Bayesian meta-analysis showed strong evidence of publication bias (BF_pb = 91.33). The model-averaged effect size was close to zero, g = 0.03 [0.00, 0.24], with evidence against an effect (BF_{10} = 0.273). Heterogeneity was strongly supported ($BF_h \longrightarrow \infty$), consistent with the frequentist analysis. Figure 5 plots posterior estimates (Cohen's *g*) alongside the prior.

Figure 5

Prior and Posterior Distributions of Effect Size (μ) for Referent Retention (A) and Referent Selection (B) from Robust Bayesian Meta-Analysis models



Note. The plot shows the posterior distribution of effect size (Cohen's g) from RoBMA models. The x-axis represents effect size, and the y-axis shows density. The posterior reflects updated beliefs after incorporating study data. Correction methods are indicated: red for PET and PEESE, blue for PEESE in RoBMA models.

Most of the bias signal arose from PET (*posterior probability* = 0.79) and PEESE (0.20); selection models contributed little. Given the strong evidence of publication bias, particularly from PET, the model-averaged estimate should be interpreted with caution. The true effect, if it exists, is likely smaller than the uncorrected frequentist estimate.

3.4.2.2 Referent Selection

For referent selection, the meta-analysis offered weak evidence for an effect (BF_{10} = 1.82), with *g* = 0.31 [0.00, 0.89]. Heterogeneity was again strongly supported ($BF_h \longrightarrow \infty$). Evidence for publication bias was also substantial (BF_pb = 77.25), with both PET and PEESE contributing (PET *posterior probability* = 0.48; PEESE = 0.51).

3.4.3 Analysis of Lewis et al. (2020)

Our sample may be biased towards significant referent selection effects, as mutual exclusivity is often a prerequisite for reporting retention data. To investigate, we reanalysed Lewis et al.'s data, focusing on typically developing children. This subset increased sample size (*n* = 101), enhancing statistical power for detecting and correcting publication bias. Additionally, as their original analyses did not account for publication bias, our approach provides further insight into the mutual exclusivity effect.

We replicated their mutual exclusivity trials, initially finding *d* = 1.37 [1.00, 1.75]. Switching to Hedges' *g* for small sample bias correction and consistency with our analysis, we obtained *g* = 1.29 [0.99, 1.59]. Using a non-nested model with paper as a random effect (yielding similar results to a nested specification), we then applied our RoBMA model.

The analysis provided overwhelming evidence of publication bias ($BF_pb > 1.6 \times 10^{11}$), far exceeding that of our referent selection analysis. The PET model dominated (*posterior probability* = 0.95), while PEESE contributed minimally (0.04), highlighting PET's strong influence on model-averaged estimates. Unlike our referent selection analysis, the Lewis et al. reanalysis found strong evidence against a meaningful effect (BF_{10} = 0.053), with an effect size near zero (g = 0.005 [0.000, 0.047]).

3.4.4 Exploratory Analyses of Publication Bias Attenuation

All these analyses consistently showed substantial heterogeneity and publication bias, with particularly striking corrections in the Lewis et al. reanalysis, where a large initial effect was

reduced to near zero. To understand this attenuation and examine robustness under different methods, we conducted additional exploratory analyses, noting no single approach is conclusive (Carter et al., 2019).

We first applied trim and fill (Duval & Tweedie, 2000), a commonly used but sometimes anti-conservative technique (Carter et al., 2019). For retention, it identified 13 missing studies, lowering the effect to $g = 0.24$ [0.15, 0.32], $p < .001$. In contrast, it estimated no missing studies for selection or the Lewis et al. reanalysis, both of which retained strong, significant effects ($g = 1.17$ [0.91, 1.43] and $g = 1.20$ [0.98, 1.42], respectively, both $p < .001$). These results suggest a more modest effect for retention but substantial effects for selection and for Lewis et al.

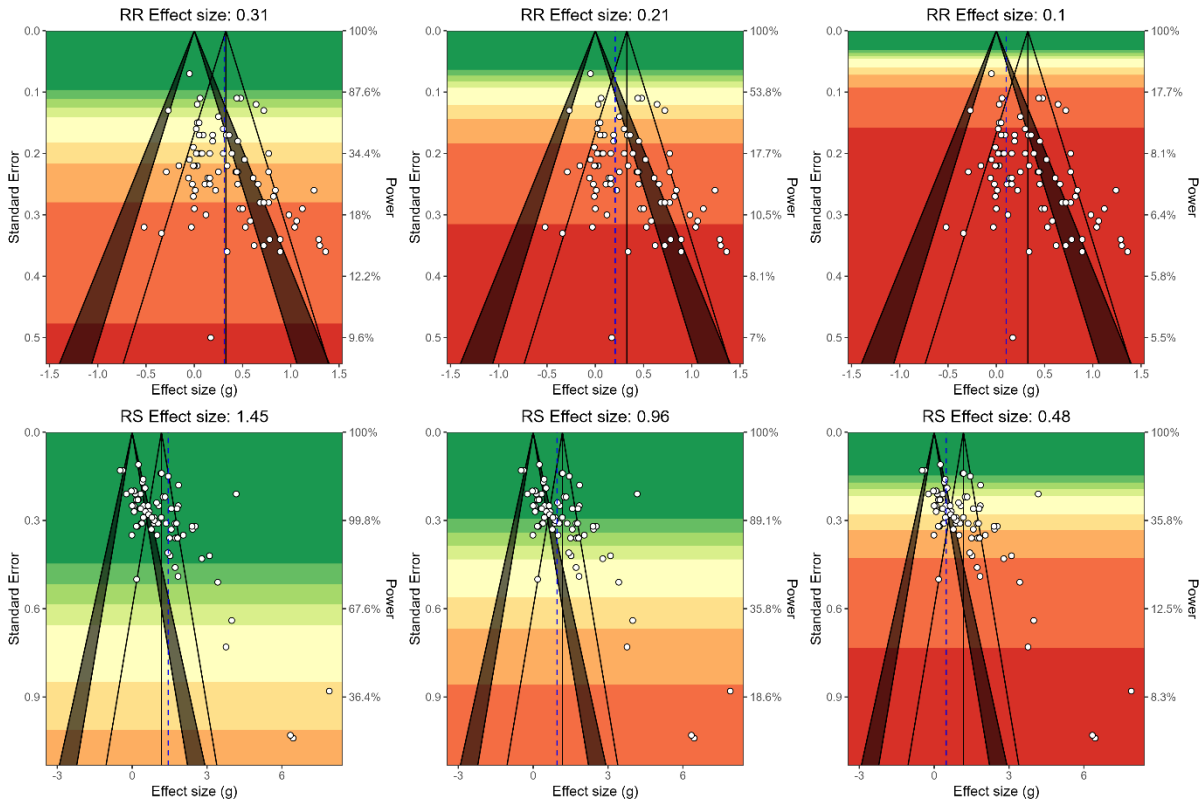
Next, to clarify the impact of PET and PEESE in RoBMA, we re-ran analyses excluding PET, which can underestimate true effects, particularly under high heterogeneity (Stanley, 2017; Stanley & Doucouliagos, 2014), and where a non-null effect is likely. The revised models yielded: for retention, $g = 0.10$ [0.00, 0.33], $BF_{10} = 1.72$ (weak evidence); for selection, $g = 0.49$ [0.00, 0.97], $BF_{10} = 16.71$ (strong evidence); for Lewis et al., $g = 0.45$ [0.24, 0.67], $BF_{10} = 796.44$ (decisive evidence). Publication bias remained robust across all datasets ($BF_pb = 25.21$, 53.3, and 1.35×10^9, respectively). These corrections generally reduced effect sizes to about one-third of the uncorrected estimates, with only the Lewis et al. dataset showing a precise nonzero effect.

3.5 Power Analysis

We conducted a power analysis to assess the statistical power of the studies included in our meta-analysis for referent retention and referent selection. Figure 6 shows the sunset plot for our observed meta-analytic effect sizes, along with more conservative estimates of 66% and 33% of the observed effect size.

Figure 6

Sunset Plots for Referent Retention (Top) and Referent Selection (Bottom) at Different Effect Sizes



Note. The plots show the statistical power of studies at different sample sizes for: original effects (left), 66% attenuation (middle) and 33% attenuation (right). Darkened contours represent p-values between .01 and .05. The x-axis represents the effect size, while the y-axis represents the standard error.

Table 2 presents the power analysis results for various effect sizes, This table shows the effect sizes, the corresponding power at the median sample size (N = 20), and the required sample size for achieving 80% power.

Table 2

Power Across Different Effect Sizes

Effect Size	Power at Median N	N for 80% Power
0.1	0.06	1571
0.2	0.09	394
0.3	0.15	176
0.4	0.23	100
0.5	0.34	64
0.6	0.46	45
0.7	0.58	34
0.8	0.69	26

Note. Median N = 20.

As can be seen in the table and accompanying figure, the power of the studies included in our meta-analysis of retention data is generally low to very low, particularly for smaller effect sizes. At the median sample size of 20 participants, the power for the most optimistic observed effect size of 0.31 is only 15%, which is far below the conventional threshold of 80%. To achieve 80% power for detecting an effect size of 0.31, a sample size of 164 would be required. In contrast, due to the much larger effect size in the referent selection

meta-analysis (which had the same median sample size), the power to detect an effect at the estimated effect size at the median sample size was much higher, but would still be low with more conservative estimates of the effect size.

4. Discussion

The present meta-analysis investigated the retention of word mappings formed through mutual exclusivity in young children. Frequentist analyses indicated a small to medium sized overall retention effect, suggesting that children may retain word–object associations established via mutual exclusivity to some extent. In contrast, our pre-registered Bayesian analyses provided some evidence against a true retention effect and suggested considerable attenuation, highlighting the potential influence of publication bias in this literature. However, our pre-registered approach to publication bias may have been overly conservative, especially given the high heterogeneity seen in our data. Further exploratory analysis offers some reconciliation with our frequentist analysis, as we discuss below. By comparison, referent selection demonstrated a larger and more reliable effect size despite also showing evidence of publication bias, aligning with prior research that shows a sizable and reliable effect of mutual exclusivity in referent selection tasks (Lewis et al., 2020).

The discrepancy between the larger effect size for referent selection and the smaller effect size for referent retention has implications for the processes underlying word learning. Our findings suggest that while mutual exclusivity is an effective strategy for disambiguating novel words, it may not inherently support the retention of these novel mappings. Neither classic accounts of mutual exclusivity which cast it as a constraint on learning (e.g., Markman, 1989, 1990; Mervis & Bertrand, 1994) nor accounts involving social-pragmatic inference (e.g., Diesendruck & Markson, 2001; Frank et al., 2009) consider longer-term learning outcomes. Instead, our results are consistent with models of word learning proposing a dissociation between referent selection and associative memory processes (McMurray et al., 2012). We also note an interesting parallel in the adult word learning literature. While there have been suggestions that mutual exclusivity may promote the addition of new word mappings to the lexicon in adults (Coutanche & Thompson-Schill, 2014), more recent converging evidence does not support this view (Gurunandan et al., 2023; Lindsay et al., 2025; O’Connor et al., 2019).

4.1 Moderators and developmental effects

In exploring potential moderators of the retention effect, we identified age as a significant predictor when analysed separately from vocabulary size, indicating developmental improvements in children's ability to retain word mappings formed through mutual exclusivity. Older children showed greater retention of these mappings than younger children, suggesting that as children develop, they become more proficient not only at selecting the correct referent but also at encoding and retaining novel word–object associations (Bion et al., 2013). When age was analysed alongside vocabulary size, which was highly correlated with age, it was no longer a significant predictor. This finding implies that vocabulary size may mediate the relationship between age and retention ability. It is possible that children with larger vocabularies have more developed learning mechanisms that facilitate the formation and retention of new word mappings (e.g., Ha et al., 2020; Werker et al., 2002). Inconsistent vocabulary measures across studies, however, limited this moderator analysis to only a subset of studies.

We further investigated task-specific moderators, such as the number of mappings and the retention interval, but found null results in each case. These null findings should be interpreted cautiously due to the low statistical power in our analyses, stemming from the small number of high-powered studies in our sample. This limitation likely contributed to the large unexplained variance often observed in meta-analyses of developmental phenomena (Bergmann & Cristia, 2016; Tsui et al., 2019). Data availability posed another significant challenge for some moderators. For instance, most studies tested retention either immediately or after short delays (e.g., five minutes), with few examining longer intervals, restricting our ability to generalise findings to long-term retention of word mappings. Future research examining retention over extended periods would provide valuable insights into the durability of word mappings formed through mutual exclusivity. The lack of clear evidence for moderator effects, except for age, impedes our ability to refine models of word learning and limits our understanding of the factors influencing mutual exclusivity in retention tasks.

4.2 Publication Bias

The meta-analysis initially suggested a moderate effect size (r = 0.3) for mutual exclusivity retention, but this estimate shrank considerably when accounting for publication bias. Our analyses revealed strong evidence of publication bias across different detection

methods, despite our efforts to include unpublished data and grey literature. This bias may stem from retention being a secondary outcome in many studies, where selective reporting biases can be more pronounced (Ioannidis et al., 1998), in this case where retention would only be reported when it was significant. The potential for greater bias in secondary outcomes highlights the importance of transparent reporting through pre-registration (Havron et al., 2020; Bergmann et al., 2018; Davis-Kean & Ellis, 2019). We also found substantial evidence for publication bias in our selection data. Our analysis extends the work of Lewis et al. (2020) who did not consider publication bias in their initial analyses, revealing exceptionally strong evidence for bias within their original dataset.

Our pre-registered use of Robust Bayesian Meta-Analysis (RoBMA) led to substantial attenuation of effect sizes in both retention and selection tasks, with retention estimates reduced to zero. Applying the same methodology to the larger Lewis et al. (2020) dataset for referent selection similarly reduced their large uncorrected effect to around zero. However, all publication bias attenuation methods struggle with high heterogeneity and small sample sizes (Carter et al., 2019), which characterise both the mutual exclusivity retention and selection literature. In our case, the attenuation is largely attributed to the dominant effects of the PET method within RoBMA, which provided the best fit but is known to be overly conservative, especially with high heterogeneity (Stanley, 2017). Importantly, the PET method may overcorrect when a true non-zero effect exists, thereby increasing the risk of false negatives. (Stanley, 2017).

Given these methodological challenges, we compared our pre-registered approach with multiple methods to assess the robustness of our conclusions. We used PEESE on its own, which has been recommended as a more accurate estimate when a true effect is likely present (Stanley, 2017). This approach differs from our pre-registered RoBMA method, which included both PET and PEESE. Our exploratory analyses using PEESE suggest an effect size of around 0.1 for retention tasks, which, while much smaller than initial estimates, still provided evidence for an effect, albeit only weak evidence. We maintain that the evidence, though attenuated by publication bias correction, supports the existence of a real, albeit small, effect in retention tasks. This is especially true for older children, given that age emerged as a significant positive moderator in our analyses.

4.3 Considerations for Developmental Research

Meta-analyses in psychology, particularly in fields like infant development, often struggle to fully account for heterogeneity across studies (Bergmann & Cristia, 2016; Tsui et al., 2019). Differences in methodologies, sample populations, and definitions of key variables complicate the generalisation of findings (Oakes, 2017). Our meta-analysis encountered significant heterogeneity arising from variations in task designs and participant characteristics. Small sample sizes further reduce the reliability of effect size estimates, producing wide confidence intervals and lowering the precision of meta-analytic results (Button et al., 2013). Our power analysis indicated that, with a median sample size of 20 participants, the power to detect the uncorrected retention effect is only 15%, which is well below the recommended 80% (Oakes, 2017). For more conservative corrected effect size estimates, this power would be even lower. This very low power increases the risk of Type II errors and may contribute to the non-significant findings for many of the moderators. Low-powered studies also contribute to greater variability in effect size estimates, exacerbating the heterogeneity observed across studies and further complicating the interpretation of our meta-analytic findings.

Our findings point to several critical improvements. High-powered studies with larger sample sizes are essential to more accurately estimate meta-analytic effects and detect potential moderators (Button et al., 2013). Greater attention to measurement precision could enhance the reliability and interpretability of individual studies (see Lindsay and Mather, 2022), thereby improving the quality of data available for meta-analyses. Pre-registration and open science practices provide important safeguards against bias and help establish the robustness of research findings (Nosek et al., 2018). Collaborative research initiatives, such as the ManyBabies projects, offer a practical approach to addressing individual study limitations (Frank et al., 2017). These large-scale collaborations promote methodological standardisation, and significantly increase sample sizes, enhancing statistical power and generalisability (Bergmann et al., 2018; Byers-Heinlein et al., 2020). Addressing methodological limitations, such as standardising measures and accounting for ceiling effects, will further improve research quality (Frank et al., 2017). These improvements will collectively strengthen the robustness and reliability of research in this field.

4.4 Conclusions

Our findings highlight some avenues for future research. If retention effects are small relative to disambiguation effects, then high-powered studies are of particular importance for understanding the memory processes involved in early word learning. Furthermore, our survey of mutual exclusivity retention studies revealed that they predominantly focus on brief retention intervals. More investigation of longer-term retention would provide greater insight into the durability of word mappings (see e.g., Axelsson et al., 2025). While we have focussed on mutual exclusivity in our meta-analysis, these points are relevant to the study of other word learning cues. Future methodological improvements will be crucial for understanding the contribution of mutual exclusivity to vocabulary development and the learning of word mappings more generally.

Data Availability Statement

The project materials that support this study are openly available in the Open Science Framework at https://osf.io/uwb69. This includes the coding sheet and coding key, scripts for data analysis, email templates for contacting authors, the Stage 1 report serving as a pre-registration, and a document detailing changes from the pre-registration.

References

Axelsson, E. L., & Horst, J. S. (2014). Contextual repetition facilitates word learning via fast mapping. *Acta Psychologica, 152,* 95–99. https://doi.org/10.1016/j.actpsy.2014.08.002

Axelsson, E. L., Horst, J. S., Playford, S. L., & Winiger, A. I. (2025). Toddlers' looking behaviours during referent selection and relationships with immediate and delayed retention. *Journal of Memory and Language, 141*, 104596. https://doi.org/10.1016/j.jml.2024.104596

Axelsson, E. L., Swinton, J., Winiger, A. I., & Horst, J. S. (2018). Napping and toddlers' memory for fast-mapped words. *First Language, 38*(6), 582–595. https://doi.org/10.1177/0142723718785490

Bartoš, F. (2025, February 4). *Hierarchical Bayesian model-averaged meta-analysis* [Vignette]. RoBMA (Version 3.4.0). https://fbartos.github.io/RoBMA/articles/HierarchicalBMA.html

Bartoš, F., Maier, M., Quintana, D. S., & Wagenmakers, E. J. (2022). Adjusting for publication bias in JASP and R: Selection models, PET-PEESE, and robust Bayesian meta-analysis. *Advances in Methods and Practices in Psychological Science, 5*(3), 1 – 19. https://doi.org/10.1177/25152459221109259

Bartoš, F., Maier, M., Wagenmakers, E. J., Doucouliagos, H., & Stanley, T. D. (2023). Robust Bayesian meta-analysis: Model-averaging across complementary publication bias adjustment methods. *Research Synthesis Methods, 14*(1), 99–116. https://doi.org/10.1002/jrsm.1594

Bergmann, C., & Cristia, A. (2016). Development of infants' segmentation of words from native speech: A meta-analytic approach. *Developmental Science, 19*(6), 901-917. https://doi.org/10.1111/desc.12341

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

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

Bleijlevens, N., & Behne, T. (2024). Young children and adults use reasoning by exclusion rather than attraction to novelty to disambiguate novel word meanings. *Developmental Psychology, 60*(7), 1313-1330. https://doi.org/10.1037/dev0001705

Bloom, P. (2000). *How children learn the meanings of words*. MIT Press.

Borovsky, A. (2020). When slowing down processing helps learning: Lexico-semantic structure supports retention, but interferes with disambiguation of novel object-label mappings. *Developmental Science, 23*(6), e12963 https://doi.org/10.1111/desc.12963

Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience, 14*(5), 365-376. https://doi.org/10.1038/nrn3475

Byers-Heinlein, K., Bergmann, C., Davies, C., Frank, M. C., Hamlin, J. K., Kline, M., ... & Soderstrom, M. (2020). Building a collaborative psychological science: Lessons learned from ManyBabies 1. *Canadian Psychology/Psychologie canadienne, 61*(4), 349-363. https://doi.org/10.31234/osf.io/dmhk2

Byers-Heinlein, K., & Werker, J. F. (2009). Monolingual, bilingual, trilingual: Infants' language experience influences the development of a word-learning heuristic. *Developmental Science, 12*(5), 815–823. https://doi.org/10.1111/j.1467-7687.2009.00902.x

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development, 15*, 17–29.

Carter, C. K., & Hartley, C. (2021). Are children with autism more likely to retain object names when learning from colour photographs or black-and-white cartoons? *Journal*

of Autism and Developmental Disorders, 51(9), 3050–3062. https://doi.org/10.1007/s10803-020-04771-2

Carter, E. C., Schönbrodt, F. D., Gervais, W. M., & Hilgard, J. (2019). Correcting for bias in psychology: A comparison of meta-analytic methods. *Advances in Methods and Practices in Psychological Science, 2*(2), 115-144. https://doi.org/10.1177/2515245919847196

Cheung, R. W., Hartley, C., & Monaghan, P. (2022). Multiple mechanisms of word learning in late-talking Children: A longitudinal study. *Journal of Speech, Language, and Hearing Research, 65*(8), 2978–2995. https://doi.org/10.1044/2022_jslhr-21-00610

Coutanche, M. N., & Thompson-Schill, S. L. (2014). Fast mapping rapidly integrates information into existing memory networks. *Journal of Experimental Psychology: General, 143*(6), 2296-2303. https://doi.org/10.1037/xge0000020

Davis-Kean, P. E., & Ellis, A. (2019). An overview of issues in infant and developmental research for the creation of robust and replicable science. *Infant Behavior and Development, 57*, 101339. https://doi.org/10.1016/j.infbeh.2019.101339

Dewar, K., & Xu, F. (2007). Do 9-month-old infants expect distinct words to refer to kinds? *Developmental Psychology, 43*(5), 1227–1238. https://doi.org/10.1037/0012-1649.43.5.1227

Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology, 37*(5), 630–641. https://doi.org/10.1037/0012-1649.37.5.630

Duval, S., & Tweedie, R. (2000). Trim and fill: a simple funnel-plot–based method of testing and adjusting for publication bias in meta-analysis. *Biometrics, 56*(2), 455-463. https://doi.org/10.1111/j.0006-341X.2000.00455.x

Esterline, K. (2020). *Sleep Development and Memory in Early Childhood* [Doctoral dissertation, University of Arizona]. UA Campus Repository. https://repository.arizona.edu/handle/10150/656759

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development, 59*(5, serial no. 242).

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual*. Baltimore, MD: Paul H. Brookes Publishing Company.

Frank, M. C., Bergelson, E., Bergmann, C., Cristia, A., Floccia, C., Gervain, J., ... & Yurovsky, D. (2017). A collaborative approach to infant research: Promoting reproducibility, best practices, and theory-building. *Infancy, 22*(4), 421-435. https://doi.org/10.1111/infa.12182

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using Speakers' Referential Intentions to Model Early Cross-Situational Word Learning. Psychological Science, 20(5), 578-585. https://doi.org/10.1111/j.1467-9280.2009.02335.x

Gambi, C., Pickering, M. J., & Rabagliati, H. (2021). Prediction error boosts retention of novel words in adults but not in children. *Cognition, 211*, 104650. https://doi.org/10.1016/j.cognition.2021.104650

Golinkoff, R. M., Hirsh-Pasek, K., Bailey, L. M., & Wenger, N. R. (1992). Children and Adults use lexical principles to learn new nouns. *Developmental Psychology, 28*(1), 99–108. https://doi.org/10.1037/0012-1649.28.1.99

Golinkoff, R. M., Mervis, C. B., & Hirsh-Pasek, K. (1994). Early object labels: The case for a developmental lexical principles framework. *Journal of Child Language*, *21*(1), 125–155. https://doi.org/10.1017/S0305000900008692

Grassmann, S., Stracke, M., & Tomasello, M. (2009). Two-year-olds exclude novel objects as potential referents of novel words based on pragmatics. *Cognition*, *112*(3), 488–493. https://doi.org/10.1016/j.cognition.2009.06.010

Gurteen, P. M., Horne, P. J., & Erjavec, M. (2011). Rapid word learning in 13- and 17-month-olds in a naturalistic two-word procedure: Looking versus reaching measures. *Journal of Experimental Child Psychology, 109*(2), 201–217. https://doi.org/10.1016/j.jecp.2010.12.001

Gurunandan, K., Cooper, E., Tibon, R., Henson, R. N., & Greve, A. (2023). No evidence of fast mapping in healthy adults using an implicit memory measure: failures to replicate the lexical competition results of Coutanche and Thompson-Schill (2014). *Memory, 31*(10), 1320-1339. https://doi.org/10.1080/09658211.2023.2262188

Ha, O. R., Cashon, C. H., Holt, N. A., & Mervis, C. B. (2020). Development of rapid word–object associations in relation to expressive vocabulary: Shared commonalities in infants and toddlers with and without Williams syndrome. *Developmental Science, 23*(6), e12966. https://doi.org/10.1111/desc.12966

Halberda, J. (2003). The development of a word-learning strategy. *Cognition, 87*, B23–B34. https://doi.org/10.1016/S0010-0277(02)00186-5

Havron, N., Bergmann, C., & Tsuji, S. (2020). Preregistration in infant research—A primer. *Infancy*, 25, 734–754. https://doi.org/10.1111/infa.12353

Hilton, M. (2016). *Temperament and early word learning: the effect of shyness on referent selection and retention* [Doctoral dissertation, Lancaster University]. Lancaster EPrints. https://eprints.lancs.ac.uk/id/eprint/82171/

Hinne M., Gronau, Q. F., van den Bergh, D., & Wagenmakers, E.-J. (2020). A conceptual introduction to Bayesian model averaging. *Advances in Methods and Practices in Psychological Science, 3*, 200–215. https://doi.org/10.1177/2515245919898657

Horn, P., Fritzsche, T., Ehlert, A., & Adani, F. (2021). Tapping into the interplay of lexical and number knowledge using fast mapping: A longitudinal eye-tracking study with two-year-olds. *Infant Behavior and Development, 64*, 101573. https://doi.org/10.1016/j.infbeh.2021.101573

Horst, J. S. (2007). *Turning novel names into known names: Understanding referent selection and retention in 24-month-old children and neural networks* (Publication No. 304858441) [Doctoral dissertation, University of Iowa]. ProQuest Dissertations & Theses Global.

Horst, J. S. & Samuelson, L. K. (2008). Fast mapping but poor retention by 24-month-old infants. *Infancy, 13*(2), 128 – 157. https://doi.org/10.1080/15250000701795598

Horst, J. S., Samuelson, L. K., Kucker, S. C., & McMurray, B. (2011). What's new? Children prefer novelty in referent selection. *Cognition, 118*(2), 234 – 244. https://doi.org/10.1016/j.cognition.2010.10.015

Horst, J.S., Scott, E.J. & Pollard, J.A. (2010). The role of competition in word learning via referent selection. *Developmental Science, 13*(5), 706-713. https://doi.org/10.1111/j.1467-7687.2009.00926.x

Houston-Price, C., Caloghiris, Z., & Raviglione, E. (2010). Language experience shapes the development of the mutual exclusivity bias. *Infancy, 15*(2), 125–150. https://doi.org/10.1111/j.1532-7078.2009.00009.x

Huang, R., Wang, T., Robertson, W., and Minhas, A. (2022). Word Learning Ability in 24-Month-Olds: Interaction of Mother's Work Status and Education Level. In Y. Gong and F. Kpogo (Eds.) *Proceedings of the 46th annual Boston University Conference on Language Development* (pp. 351 - 363). Somerville, MA: Cascadilla Press.

Ioannidis, J. P., Cappelleri, J. C., & Lau, J. (1998). Issues in comparisons between meta-analyses and large trials. *JAMA, 279*(14), 1089-1093. https://doi.org/10.1001/jama.279.14.1089

Kalashnikova, M., Escudero, P., & Kidd, E. (2018). The development of fast-mapping and novel word retention strategies in monolingual and bilingual infants. *Developmental Science, 21*, e12674. https://doi.org/10.1111/desc.12674

Kucker, S. C. (2023a). *Gallagher and Kucker (unpublished)* [Unpublished raw data]. Southern Methodist University.

Kucker, S. C. (2023b). *Kucker and Chmielewski (in preparation)* [Unpublished raw data]. Southern Methodist University.

Kucker, S. C. (2023c). *Kucker and Kaipa (in preparation)* [Unpublished raw data]. Southern Methodist University.

Kucker, S. C., Mcmurray, B., & Samuelson, L. K. (2015). Slowing down fast mapping: Redefining the dynamics of word learning. *Child Development Perspectives, 9*(2), 74–78. https://doi.org/10.1111/cdep.12110

Kucker, S.C., McMurray, B. and Samuelson, L.K. (2018). Too much of a good thing: How novelty biases and vocabulary influence known and novel referent selection in 18-month-old children and associative learning models. *Cognitive Science, 42*, 463-493. https://doi.org/10.1111/cogs.12610

Kucker, S. C., McMurray, B., & Samuelson, L. K. (2020). Sometimes it is better to know less: How known words influence referent selection and retention in 18- to 24-month-old children. *Journal of Experimental Child Psychology*, *189*(104705). https://doi.org/10.1016/j.jecp.2019.104705

Kucker, S., McMurray, B., & Samuelson, L. (2024). Examining the incremental process of word learning: Word-form exposure and retention of new word-referent mappings. *Language Development Research,* 4(1), 326-360. doi: 10.34842/3ppz-g170

Kucker, S. C., & Samuelson, L. K. (2012). The First Slow Step: Differential Effects of Object and Word-Form Familiarization on Retention of Fast-Mapped Words. *Infancy, 17*(3), 295–323. https://doi.org/10.1111/j.1532-7078.2011.00081.x

Kucker, S. C., & Seidler, E. (2023). The timescales of word learning in children with language delays: In-the-moment mapping, retention, and generalization. *Journal of Child Language, 50*(2), 245–273. https://doi.org/10.1017/s0305000921000817

Lewis, M., Cristiano, V., Lake, B. M., Kwan, T., & Frank, M. C. (2020). The role of developmental change and linguistic experience in the mutual exclusivity effect. *Cognition, 198,* 104191. https://doi.org/10.1016/j.cognition.2020.104191

Lindsay, S., Lucas, A., O'Connor, R., Riggs, K., & Gaskell, G. (2025). *Fast mapping does not enhance lexical integration for spoken or written novel words*. Manuscript in preparation.

Lindsay, S., & Mather, E. (2022). Developmental psychologists should care about measurement precision. *Infant and Child Development, 31*(5), e2321. https://doi.org/10.1002/icd.2321

Maier, M., Bartoš, F., & Wagenmakers, E. J. (2023). Robust Bayesian meta-analysis: Addressing publication bias with model-averaging. *Psychological Methods, 28*(1), 107. https://doi.org/10.1037/met0000405

Markman, E. M. (1989). *Categorization and naming in children: Problems of induction.* Cambridge, MA: MIT Press.

Markman, E. M. (1990). Constraints children place on word meanings. *Cognitive Science, 14*(1), 57 – 77. https://doi.org/10.1016/0364-0213(90)90026-S

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

Markson, L., & Bloom, P. (1997). Evidence against a dedicated system for word learning in children. *Nature, 385*(6619), 813–815. https://doi.org/10.1038/385813a0

Mather, E., & Plunkett, K. (2010). Novel labels support 10-month-olds' attention to novel objects. *Journal of Experimental Child Psychology, 105*(3), 232–242. https://doi.org/10.1016/j.jecp.2009.11.004

Mather, E., & Plunkett, K. (2011). Mutual exclusivity and phonological novelty constrain word learning at 16 months. *Journal of Child Language, 38*(5), 933–950. https://doi.org/10.1017/s0305000910000401

Mather, E., & Plunkett, K. (2012). The role of novelty in early word learning. *Cognitive Science, 36*(7), 1157–1177. https://doi.org/10.1111/j.1551-6709.2012.01239.x

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review, 119*(4), 831–877. https://doi.org/10.1037/a0029872

Mervis, C. B., & Bertrand, J. (1994). Acquisition of the novel name-nameless category (n3c) principle. *Child Development, 65*(6), 1646–1662. https://doi.org/10.2307/1131285

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G. (2009). Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ, 339*, b2535. https://doi.org/10.1136/bmj.b2535

Morin-Lessard, E. (2018). *Speech Disfluencies: Their Role in Comprehension and Word Learning* [Doctoral dissertation, Concordia University]. Spectrum: Concordia

University Research
Repository. https://spectrum.library.concordia.ca/id/eprint/984575/

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences, 115*(11), 2600-2606. https://doi.org/10.1073/pnas.1708274114

Oakes, L.M. (2017), Sample size, statistical power, and false conclusions in infant looking-time research. *Infancy, 22,* 436-469. https://doi.org/10.1111/infa.12186

O'Connor, R. J., Lindsay, S., Mather, E., & Riggs, K. J. (2019). Why would a special FM process exist in adults, when it does not appear to exist in children? *Cognitive Neuroscience, 10*(4), 221-222. https://doi.org/10.1080/17588928.2019.1574260

Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews, 5*(1), 1-10. https://doi.org/10.1186/s13643-016-0384-4

Pino Escobar, G. (2021). *Word learning and executive functions in preschool children : bridging the gap between vocabulary acquisition and domain-general cognitive processes.* [Doctoral dissertation, Western Sydney University]. Western Sydney University Research Profiles and Repository. http://hdl.handle.net/1959.7/uws:67860

Pomper, R. (2020). *Combining Cues in Novel Word Learning.* [Doctoral dissertation, University of Wisconsin-Madison]. UW Electronic Theses and Dissertations. https://asset.library.wisc.edu/1711.dl/2GXW7LBCIHQMT8B/R/file-6ca40.pdf

Pomper, R. & Saffran, J. R. (2019). Familiar object salience affects novel word learning. *Child Development, 90,* 246-262. https://doi.org/10.1111/cdev.13053

Reuter, T., Borovsky, A., & Lew-Williams, C. (2019). Predict and redirect: Prediction errors support children's word learning. *Developmental Psychology, 55*(8), 1656–1665. https://doi.org/10.1037/dev0000754

Rocha-Hidalgo, J., Feller, M., Blanchfield, O. A., Kucker, S. C., & Barr, R. F. (2021). Patterns of mutual exclusivity and retention: A study of monolingual and bilingual 2-year-olds. *Infancy, 26*(6), 1011–1036. https://doi.org/10.1111/infa.12432

Quine, W. V. O. (1960). *Word and Object*. MIT Press.

Schafer, G. & Plunkett, K. (1998). Rapid word learning by fifteen-month-olds under tightly controlled conditions. *Child Development, 69*(2), 309–320. https://doi.org/10.1111/j.1467-8624.1998.tb06190.x

Schwarzer, G., Carpenter, J. R., & Rücker, G. (2015). *Meta-analysis with R*. Springer.

Sia, M. Y., & Mayor, J. (2022). Improvements of statistical learning skills allow older children to go beyond single-hypothesis testing when learning words. *Journal of Child Language, 49*(6), 1268–1280. https://doi.org/10.1017/s0305000921000532

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568. https://doi.org/10.1016/j.cognition.2007.06.010

Stanley, T. D. (2017). Limitations of PET-PEESE and other meta-analysis methods. *Social Psychological and Personality Science, 8*(5), 581-591. https://doi.org/10.1177/1948550617693062

Stanley, T. D., & Doucouliagos, H. (2014). Meta-regression approximations to reduce publication selection bias. *Research Synthesis Methods*, *5*(1), 60-78. https://doi.org/10.1002/jrsm.1095

Tsui, A. S. M., Byers-Heinlein, K., & Fennell, C. T. (2019). Associative word learning in infancy: A meta-analysis of the switch task. *Developmental Psychology, 55*(5), 934 - 950. https://doi.org/10.1037/dev0000699

Twomey, K. E., Ranson, S. L., & Horst, J. S. (2014). That's more like it: multiple exemplars facilitate word learning. *Infant and Child Development*, *23*(2), 105–122. https://doi.org/10.1002/icd.1824

Twomey, K. E., Ma, L., & Westermann, G. (2018). All the right noises: Background variability helps early word learning. *Cognitive Science, 42*(S2), 413–438. https://doi.org/10.1111/cogs.12539

van Erp, S., Verhagen, J., Grasman, R. P., & Wagenmakers, E. J. (2017). Estimates of between-study heterogeneity for 705 meta-analyses reported in Psychological Bulletin

from 1990–2013. *Journal of Open Psychology Data*, *5*(1), Article 4. https://doi.org/10.5334/jopd.33

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

Vlach, H. A. (2019). Learning to remember words: Memory constraints as double-edged sword mechanisms of language development. *Child Development Perspectives*, *13*(3), 159–165. https://doi.org/10.1111/cdep.12337

Walters, W. H. (2007). Google Scholar coverage of a multidisciplinary field. *Information Processing & Management*, *43*(4), 1121-1132. https://doi.org/10.1016/j.ipm.2006.08.006

Weatherhead, D., Arredondo, M. M., Nácar Garcia, L., & Werker, J. F. (2021). The Role of Audiovisual Speech in Fast-Mapping and Novel Word Retention in Monolingual and Bilingual 24-Month-Olds. *Brain Sciences*, *11*(1), 114. https://doi.org/10.3390/brainsci11010114

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: Effects of age and vocabulary size. *Infancy*, *3*(1), 1-30. https://doi.org/10.1207/S15327078IN0301_1

Wilkinson, K. M., Ross, E., & Diamond, A. (2003). Fast mapping of multiple words: Insights into when “the information provided” does and does not equal “the information perceived.” *Journal of Applied Developmental Psychology*, *24*(6), 739–762. https://doi.org/10.1016/j.appdev.2003.09.006

Yurovsky, D., & Frank, M. C. (2017). Beyond naïve cue combination: salience and social cues in early word learning. *Developmental Science*, *20*(2), e12349. https://doi.org/10.1111/desc.12349

Table 1

Variable	Mean	Median	SD	Min	Max
Sample size (RR task)	27.7	20.0	26.7	4.0	209.0
Sample size (RS task)	28.3	20.0	27.5	4.0	220.0
Age (months)	31.2	29.3	10.8	16.3	66.1
Productive vocabulary size (MCDI-WS)	261.7	270.7	139.0	28.0	472.0
Number of novel mappings	3.4	3.0	1.9	1.0	8.0
Number of retention trials per novel mapping	2.6	2.0	1.8	1.0	6.0
Number of familiar distractors in selection task	1.4	1.0	0.5	1.0	2.0
Number of selection trials per familiar distractor	1.4	1.0	1.4	0.0	6.0
Total number of objects in retention task	1.6	1.0	0.7	1.0	3.0
Retention interval (minutes)	106.9	0.0	366.7	0.0	1440.0

Table 2

Effect Size	Power at Median N	N for 80% Power
0.1	0.06	1571
0.2	0.09	394
0.3	0.15	176
0.4	0.23	100
0.5	0.34	64
0.6	0.46	45
0.7	0.58	34
0.8	0.69	26

Appendix

Table A1

Summary of Studies Using Intermodal Preferential Looking (IPL) Paradigm

Study	Condition	N	Age	g RR	g RS	SE RR	SE RS
Bion et al. (2013)	18-Month-Olds	22	18.8	-0.05	0.13	0.21	0.21
Bion et al. (2013)	24-Month-Olds	25	24.9	0.06	1.07	0.2	0.25
Bion et al. (2013)	30-Month-Olds	20	30.8	0.45	1.17	0.23	0.29
Borovsky (2020)	High Density Condition	65	24.8	0.03	-0.38	0.12	0.13
Borovsky (2020)	Low Density Condition	65	24.8	-0.27	-0.48	0.13	0.13
Horn et al. (2021)	30-Month-Olds	25	30.1	0.3	0.98	0.2	0.24
Horn et al. (2021)	36-Month-Olds	25	36.4	0.77	1.72	0.23	0.31
Huang et al. (2022)	Lme Working	8	24.5	0.34	-0.01	0.36	0.35
Huang et al. (2022)	Lme At Home	11	24.5	-0.52	0.36	0.32	0.31
Huang et al. (2022)	Mme Working	20	24.5	0.25	0.25	0.23	0.23

Huang et al. (2022)	Mme At Home	17	24.5	0.45	0.69	0.25	0.27
Huang et al. (2022)	Hme Working	16	24.5	0.68	0.32	0.28	0.26
Huang et al. (2022)	Hme At Home	4	24.5	0.17	0.18	0.5	0.5
Kalashnikova et al. (2018)	Expt 1 Monolingual	33	18.3	0.77	0.52	0.2	0.19
Kalashnikova et al. (2018)	Expt 1 Bilingual	25	18.3	0.52	0.34	0.21	0.21
Kalashnikova et al. (2018)	Expt 2 Monolingual	18	24.5	0.82	0.46	0.27	0.25
Kalashnikova et al. (2018)	Expt 2 Bilingual	13	24.5	0.47	0.88	0.29	0.31
Mather & Plunkett (2011)	Familiar Phonological Neighbour (Label = Pok)	21	16.3	-0.16	-0.23	0.22	0.21
Mather & Plunkett (2011)	No Familiar Phonological Neighbour (Label = Meb)	21	16.3	0.44	-0.03	0.23	0.2
Morin-Lessard (2019)	Monolingual, Fluent Condition, Block 1	16	32.3	0.98	1.04	0.3	0.31
Morin-Lessard (2019)	Monolingual, Fluent Condition, Block 2	16	32.3	0.77	0.88	0.28	0.29

Morin-Lessard (2019)	Monolingual, Disfluent Condition, Block 1	16	32.3	1.29	0.66	0.34	0.28
Morin-Lessard (2019)	Monolingual, Disfluent Condition, Block 2	15	32.3	1.36	0.88	0.36	0.3
Morin-Lessard (2019)	Bilingual, Fluent Condition, Block 1	16	32.4	1.06	0.34	0.31	0.26
Morin-Lessard (2019)	Bilingual, Fluent Condition, Block 2	13	32.4	0.5	0.45	0.29	0.29
Morin-Lessard (2019)	Bilingual, Disfluent Condition, Block 1	16	32.4	0.72	0.79	0.28	0.29
Morin-Lessard (2019)	Bilingual, Disfluent Condition, Block 2	11	32.4	0.89	0.39	0.36	0.31
Pomper & Saffran (2019)	High Salience Condition	36	40	0.09	1.24	0.17	0.22
Pomper & Saffran (2019)	Low Salience Condition	36	40	0.33	2.42	0.17	0.33
Pomper (2020)	Study 1, Me Cue	51	39.7	0.25		0.14	0.14
Reuter et al. (2019)	Constrained Condition, 3-Year-Olds	16	40	-0.02	0.07	0.25	0.25

Reuter et al. (2019)	Constrained Condition, 4-Year-Olds	20	54.2	0.03	0.64	0.22	0.25
Reuter et al. (2019)	Constrained Condition, 5-Year-Olds	18	65.6	-0.06	0.17	0.24	0.23
Reuter et al. (2019)	Unconstrained Condition, 3-Year-Olds	15	40	0.01	0.53	0.26	0.27
Reuter et al. (2019)	Unconstrained Condition, 4-Year-Olds	18	54.2	0.14	0.23	0.24	0.23
Reuter et al. (2019)	Unconstrained Condition, 5-Year-Olds	18	65.6	0.16	1.36	0.24	0.31
Twomey et al. (2018)	Constant Colour	14	22.8	-0.01	0.08	0.27	0.27
Twomey et al. (2018)	Variable Colour	16	22.8	0.16	0	0.25	0.25
Weatherhead et al. (2021)	Expt 1 Monolinguals	24	24.1	0.03	0.08	0.2	0.2
Weatherhead et al. (2021)	Expt 1 Bilinguals	24	24.1	0.04	0.09	0.2	0.2
Weatherhead et al. (2021)	Expt 2 Monolinguals	24	24.1	0.07	0.08	0.2	0.2
Weatherhead et al. (2021)	Expt 2 Bilinguals	24	24	0.08	0.09	0.2	0.2

Weatherhead et al. (2021)	Expt 3 Bilinguals	24	24	0.02	0.08	0.2	0.2

Note. 'g RR' and 'g RS' represent the Hedge’s g effect sizes for Referent Retention and Referent Selection tasks, respectively, 'SE' denotes the standard error of the effect size estimate.

Table A2

Summary of Studies Using Object Selection Paradigm

Study	Condition	N	Age	g RR	g RS	SE RR	SE RS
Axelsson et al. (2018)	Nap Group, Immediate Test	20	29.3	0.66	1.74	0.25	0.36
Axelsson et al. (2018)	Nap Group, 4 Hr Test	20	29.3	0.84	1.74	0.26	0.36
Axelsson et al. (2018)	Nap Group, Next Day Test	20	29.3	0.43	1.74	0.23	0.36
Axelsson et al. (2018)	Wake Group, Immediate Test	20	29.3	0.61	1.82	0.24	0.36
Axelsson et al. (2018)	Wake Group, 4 Hr Test	20	29.3	0	1.82	0.22	0.36
Axelsson et al. (2018)	Wake Group, Next Day Test	20	29.3	-0.29	1.82	0.23	0.36
Axelsson, Horst (2014)	Novel Repeat Comp, Known Repeat Comp	12	36.4	0.89	1.52	0.34	0.42
Axelsson, Horst (2014)	Novel Repeat Competitors, Known Competitors Not Repeat	12	36.4	0.89	1.72	0.34	0.46
Bleijlevens & Behne (2024)		33	32.9	0.45	2.4	0.18	0.32
Carter & Hartley (2021)	Photograph	20	41.5	-0.01	6.45	0.22	1.04

Carter & Hartley (2021)	Cartoon	20	41.5	1.12	6.34	0.29	1.03
Cheung et al. (2022)	Typically Developing 3-Year-Olds (T2)	36	38	0.05	0.41	0.17	0.17
Esterline (2020)	Nap Condition, Ee First	10	30.2	-0.34	0.21	0.33	0.32
Esterline (2020)	Nap Condition, Fm First	10	30.2	0.62	0.21	0.35	0.32
Esterline (2020)	Wakefulness Condition, Ee First	10	30.2	-0.03	0.17	0.32	0.32
Esterline (2020)	Wakefulness Condition, Fm First	10	30.2	0.72	0.17	0.35	0.32
Kucker (2023a)		37	28.2	0.36	1.8	0.17	0.26
Gambi et al. (2021)	Expt 6, High Constraint	80	38	0.72	1.15	0.13	0.14
Gambi et al. (2021)	Expt 6, Low Constraint	80	38	0.64	1.86	0.12	0.18
Gambi et al. (2021)	Expt 7, Implausible distractor	86	38	0.48	1.18	0.11	0.14
Gambi et al. (2021)	Expt 7, Plausible distractor	86	38	0.44	1.45	0.11	0.15
Hilton (2016)	Less-Shy Children	11	29.2	0.12	1.43	0.3	0.41
Hilton (2016)	Shy Children	12	29.2	0.58	1	0.31	0.35

Horst & Samuelson (2008)	Referent Selection And Retention Task	16	24.7	0.11	1.5	0.25	0.36
Horst (2007)	Neutral Condition Only	12	24.5	0	0.77	0.29	0.33
Kucker (2023b)		209	27.8	-0.05	4.17	0.07	0.21
Kucker (2023c)		22	30	0.34	3.99	0.22	0.64
Kucker & Samuelson (2012)		11	24.4	0.78	1.57	0.34	0.33
Kucker & Seidler (2023)	Nl Age Match	26	25.8	0.39	3.43	0.2	0.51
Kucker & Seidler (2023)	Nl Vocab Match	15	25.4	0.22	3.76	0.26	0.73
Kucker et al. (2018)	Expt 1	29	18.8	-0.01	1.57	0.19	0.26
Kucker et al. (2018)	Expt 2	26	18.5	0.16	1.78	0.2	0.31
Kucker et al. (2020)	Expt 1, Strongly Known	42	22.9	0.05	7.9	0.15	0.88
Kucker et al. (2020)	Expt 1, Weakly Known	42	22.9	0.3	1.85	0.16	0.25
Kucker et al. (2020)	Expt 2, Strongly Known	40	21.6	0.02	2.53	0.16	0.32
Kucker et al. (2020)	Expt 2, Weakly Known	37	21.6	0.19	1.32	0.17	0.22
Kucker et al. (2024)	18-Month-Olds	31	18.9	0.19	3.1	0.18	0.42

Kucker et al. (2024)	24-Month-Olds	26	24.6	0.52	2.05	0.21	0.35
Pino Escobar (2021)	Immediate Retention	15	54	1.04		0.32	0.26
Pino Escobar (2021)	Delayed Retention	15	54	1.3		0.35	0.26
Rocha-Hidalgo et al. (2021)	Monolingual	81	25.4	0.06	0.25	0.11	0.11
Rocha-Hidalgo et al. (2021)	Bilingual	44	25.7	0.03	0.44	0.15	0.16
Sia & Mayor (2022)	Four-Year-Olds	11	54.5	0.53	1.84	0.32	0.49
Sia & Mayor (2022)	Five-Year-Olds	26	66.1	1.24	2.79	0.26	0.4

Note. 'g RR' and 'g RS' represent the Hedge’s g effect sizes for Referent Retention and Referent Selection tasks, respectively, 'SE' denotes the standard error of the effect size estimate.