Journal of Experimental Child Psychology 221 (2022) 105449

Contents lists available at ScienceDirect

Journal of Experimental Child Psychology

journal homepage: www.elsevier.com/locate/jecp

Children's knowledge of superordinate words predicts subsequent inductive reasoning

Ellise Suffill [a,b,c,*], Christina Schonberg [a,b], Haley A. Vlach [a], Gary Lupyan [b]

[a] Department of Educational Psychology, University of Wisconsin–Madison, Madison, WI 53706, USA
[b] Department of Psychology, University of Wisconsin–Madison, Madison, WI 53706, USA
[c] Department of Psychology, University of Vienna, Vienna, 1010, Austria

ARTICLE INFO

Article history:
Received 27 July 2021
Revised 8 April 2022
Available online 10 May 2022

Keywords:
Word learning
Vocabulary
Language development
Superordinates
Hypernymy
Inductive reasoning

ABSTRACT

Children's early language knowledge—typically assessed using standardized word comprehension tests or through parental reports—has been positively linked to a variety of later outcomes, from reasoning tests to academic performance to income and health. To better understand the mechanisms behind these links, we examined whether knowledge of certain "seed words"—words with high inductive potential—is positively associated with inductive reasoning. This hypothesis stems from prior work on the effects of language on categorization suggesting that certain words may be important for helping people to deploy categorical hypotheses. Using a longitudinal design, we assessed 36 2- to 4-year-old children's knowledge of 333 words of varying levels of generality (e.g., *toy* vs. *pinwheel*, *number* vs. *five*). We predicted that adjusting for overall vocabulary, knowledge of more general words (e.g., *toy*, *number*) would predict children's performance on inductive reasoning tasks administered 6 months later (i.e., a subset of the Stanford–Binet Intelligence Scales for Early Childhood–Fifth Edition [SB-5] and Woodcock–Johnson Tests of Cognitive Abilities [WJ] concept formation tasks). This prediction was confirmed for one of the measures of inductive reasoning (i.e., the SB-5 but not the WJ) and notably for the task considered to be less reliant on language. Although our experimental design demonstrates only a correlational relationship between seed word knowledge and inductive reasoning ability, our results are consistent with the

* Corresponding author at: Department of Psychology, University of Vienna, Vienna, 1010, Austria.
E-mail address: ellise.suffill@univie.ac.at (E. Suffill).

https://doi.org/10.1016/j.jecp.2022.105449
0022-0965/© 2022 Elsevier Inc. All rights reserved.

possibility that early knowledge of certain seed words facilitates performance on putatively nonverbal reasoning tasks.
© 2022 Elsevier Inc. All rights reserved.

Introduction

What is the role of language in cognitive development? One way to answer this question is to examine whether differences in children's linguistic knowledge predict differences in performance on putatively nonverbal cognitive tasks. One does not need to look hard to find studies linking children's language—most often measured in terms of expressive vocabulary size—to a variety of cognitive outcomes. For example, a larger vocabulary at 16 to 24 months of age predicts performance on other language tasks, such as reading, several years later (Duff, Reen, Plunkett, & Nation, 2015). It also predicts mathematics achievement even when adjusted for a wide variety of health, sociodemographic and cognitive factors (Morgan, Farkas, Hillemeier, Hammer, & Maczuga, 2015; see also LeFevre et al., 2010; Peng et al., 2020, for a meta-analysis). Early language skills have likewise been linked to executive function (Jones et al., 2020; Kuhn, Willoughby, Vernon-Feagans, & Blair, 2016), analogical reasoning (Edwards, Figueras, Mellanby, & Langdon, 2011; Socher, Ingebrand, Wass, & Lyxell, 2020), and the understanding of false beliefs (see meta-analysis by Milligan, Astington, & Dack, 2007).

One possibility is that these positive relationships between language and cognitive outcomes simply affirm that children who are good at one thing (e.g., learning words) tend to be good at other things (e.g., inductive reasoning). That is, the positive correlations between language and cognitive development may simply reflect the well-known positive manifold of human cognition (Jensen, 1998). For example, in Morgan et al.'s (2015) analysis of more than 8000 2-year-old children, oral vocabulary size at 24 months (measured through a parental word checklist) was strongly correlated with a measure of general cognitive function assessed at the same time—a relationship that held after controlling for multiple demographic and health-related factors.

However, there are reasons to question the positive manifold explanation in favor of a mutualistic account (Kievit, 2020; Kievit et al., 2017; van der Maas et al., 2006). According to mutualistic accounts, the positive correlations in performance on various tasks emerge due to mutualistic causal influences rather than from having one common cause. One source of support for the mutualistic account is that children's language development is not simply a function of their intelligence but rather is also strongly linked to environmental factors such as the amount and quality of language that children experience (Huttenlocher, Waterfall, Vasilyeva, Vevea, & Hedges, 2010; Newman, Rowe, & Ratner, 2016; Song, Demuth, & Morgan, 2018). In some cases, differences in language environment are clearly due to entirely extrinsic reasons such as congenital deafness—a condition obviously not caused by any difference in children's cognitive abilities. Children who are born deaf and fitted with cochlear implants have poorer performance on nonverbal analogical/inductive reasoning compared with similarly aged hearing children. However, this performance difference disappears when children are matched according to language skills (Socher et al., 2020). This pattern is hard to reconcile with the idea that differences in both language and cognitive outcomes are caused by a common factor (e.g., general intelligence) but is consistent with the idea of causal links between language and cognitive development.

Further support for the mutualistic account comes from studies that found language skills to predict subsequently measured outcomes (both verbal and nonverbal) better than the reverse. Finding that nonverbal measures at Time 1 are a poor predictor of language skills at Time 2 is surprising if both are simply outcomes of general intelligence. For example, Gathercole, Willis, Emslie, and Baddeley (1992) found that vocabulary of preschoolers was more correlated with later performance on tests measuring nonverbal intelligence than the reverse. Jones et al. (2020) found that vocabulary knowledge of 8-year-olds (measured through picture naming) predicted their performance on executive function tasks assessing inhibition and switching costs 2 years later, but executive function of

8-year-olds did not predict their later vocabulary. Examining older children, Ritchie, Bates, and Plomin (2015) found that differences in language skills (i.e., reading proficiency and vocabulary) between 10-year-old identical twins was a better predictor of their differences in nonverbal reasoning at 12 years of age than the reverse. In work directly aimed to contrast mutualism with a positive manifold account, Kievit et al. (2017) found that greater vocabulary knowledge in 14- to 25-year-olds predicted greater performance on a nonverbal matrix reasoning task 1.5 years later to a stronger degree than the reverse, a finding that was subsequently replicated with 6- to 8-year-old children (Kievit, Hofman, & Nation, 2019). Taken together, these studies are consistent with the existence of a causal link from early language skills to later performance on nonverbal assessments.

Lastly, a rich experimental literature shows that language affects performance in a variety of nonverbal tasks, which supports a mutualistic account (see Lupyan, 2016, for review). Using standard experimental paradigms allows for much stronger claims than observational studies of the sort reviewed above. There is substantial evidence that labels facilitate category learning (Balaban & Waxman, 1997; Fulkerson & Waxman, 2007; Lupyan, Rakison, & McClelland, 2007; Nazzi & Gopnik, 2001; Perry & Samuelson, 2013; Plunkett, Hu, & Cohen, 2008) and promote inductive inferences (Deng & Sloutsky, 2013; Fulkerson & Waxman, 2007; Gelman & Davidson, 2013; Graham, Booth, & Waxman, 2012; Sloutsky, Lo, & Fisher, 2001). Once a category is learned, it helps to selectively activate category-diagnostic features (Edmiston & Lupyan, 2015; Lupyan & Thompson-Schill, 2012); categories with more nameable constituents are induced more easily than formally equivalent categories with less nameable features (Zettersten & Lupyan, 2020)—further evidence that category labels play an active role in seemingly nonverbal tasks. Moreover, naming impairments, such as aphasia, produce categorization impairments (Gainotti, 2014; Lupyan & Mirman, 2013), and interfering with language in healthy adults, also impairs categorization (Lupyan, 2009).

Taken together, these studies show causal links between lab-administered manipulations, such as explicit labeling of objects, and cognitive outcomes, such as categorization and category induction. However, these studies do little to advance our understanding of how learning some aspect of language during normal development contributes (or not) to the kind of cognitive skills examined by the observational studies reviewed above. In the current study, we combined the main strength of observational studies—their ecological validity—with the theoretical insights gained from experimental investigations of the links between language and cognition. We did this by examining the link between parental reports of children's word knowledge (observational) and children's performance on common tests of inductive reasoning (experimental).

A common feature of the observational studies reviewed above is their reliance on standardized language assessments, such as parental word checklists (e.g., the MacArthur–Bates Communicative Development Inventories [MCDI]; Fenson et al., Frank, Braginsky, Yurovsky, & Marchman, 2017), and picture naming tests, such as the Peabody Picture Vocabulary Test (PPVT; Dunn & Dunn, 2007), as measures of children's language development. These tests have been carefully designed to have good psychometric properties but are not designed for assessing *what* a child knows. For example, as typically used, the MCDI outcome variable is the number of words a child comprehends and/or produces, and it is this summed score that is used for correlating with cognitive outcomes. This reliance on sums without regard for what exactly the child knows (i.e., what words make up the summed score) makes it extremely difficult to understand the mechanisms behind the language–cognition links during development.

Here, we took a first step toward understanding one of the mechanisms by examining whether knowing specific *types* of words is associated with better performance on a common type of nonverbal reasoning problem—inductive reasoning (for visual patterns). Specifically, we hypothesized that knowledge of superordinate words—for reasons we describe below—may be especially useful in inductive reasoning. We are not the first to ask whether knowledge of certain words is linked to cognitive outcomes. For example, Vanluydt, Supply, Verschaffel, and Van Dooren (2021) found that children who knew the word *double* were better at solving proportion-based problems (adjusting for socioeconomic status [SES] and general vocabulary knowledge). Moreover, Miller, Vlach, and Simmering (2017) found that children's production of spatial words predicts their performance on spatial cognition tasks. Finally, Simms and Gentner (2019) investigated whether children's encoding of the midpoint, a complex spatial relation, was predicted by their knowledge of the relevant spatial

terms *middle* and *between*. Children's knowledge of the words *middle* and *between* indeed predicted their search success beyond what was predicted by age or knowledge of other spatial terms. Here, we examined the claim that knowing certain *types* of words may help children to reason in a domain that is not obviously related to the meaning of the individual words.

A link between superordinate words and inductive reasoning?

Using a word appropriately requires knowing the limits of its extension—what is and is not denoted by the word. Although all content words require distinguishing category members from nonmembers, categories denoted by some words are more heterogeneous and/or abstract than categories denoted by other words. For example, referents of a word like *dog* are much more similar to one another than referents of words like *fish* and *animal* (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Yu, Maxfield, & Zelinsky, 2016). Similar differences in generality can be seen in verbs; compare *sweep* with its characteristic motion and the use of typical instruments such as a broom, and compare *clean* with its much wider extension—sweeping, vacuuming, scrubbing, washing, and the like—a category of actions of widely varying durations and instruments, held together by something like a common result.

A key finding of Rosch et al.'s (1976) classic work is that superordinate (i.e., more semantically general) words are relatively more difficult to learn than "basic-level" words with narrower extensions (see also Mervis & Crisafi, 1982). What is difficult about learning superordinate words? One of the challenges is learning to ignore (i.e., abstract over) large and often salient perceptual differences between individual referents of the superordinate term and treating them as members of a more general category (Fenson, Cameron, & Kennedy, 1988). For example, learning the word *animal* requires treating very diverse entities such as spiders and dogs as the same kind of thing despite large (and highly noticeable) differences in their size, dietary habits, and number of legs. In contrast, words like *dog* and *sweep* tend to pick out categories whose members (individual dogs and individual acts of sweeping) already cohere based on their perceptual properties.

As discussed so far, the relationship between word learning and categorization may seem one-directional, with children mapping words onto preexisting categories (Gleitman & Fisher, 2005; Snedeker & Gleitman, 2004). However, there is reason to think that the relationship between word learning and category learning is bidirectional. Although learning a category certainly does not require first learning its name, learning labels can help children (and adults) to learn categories (e.g., Balaban & Waxman, 1997; Casasola, 2005a, 2005b; Graham et al., 2012; Lupyan et al., 2007; Nazzi & Gopnik, 2001; Plunkett et al., 2008; Waxman, 2003; Zettersten & Lupyan, 2020). This may be because the need to produce and comprehend these words provides children with increased practice in treating unlike items as similar by virtue of their shared label (a form of structural alignment; e.g., Christie & Gentner, 2010; Gentner & Namy, 1999) or because labels help to set up attractors in conceptual space (Clark & Karmiloff-Smith, 1993; Lupyan, 2012; Lupyan et al., 2007).

To the extent that learning more general (superordinate) words requires abstracting over salient differences, learning and using such words may promote learning and reasoning about higher-order relations. For example, learning a word like *color* compared with the names of specific colors may facilitate selectively attending to colors and to the relationship among them (e.g., which of these are similar *in color;* Davidoff & Roberson, 2004; Lupyan, 2009). In aggregate, a child whose vocabulary includes more general words may be better able to induce general patterns from specific objects—precisely the kind of skill tapped by tests of inductive reasoning. This link may arise either because learning more general words is a sign of a more developed ability to abstract (i.e., both have a common cause), because knowledge of specific abstract words helps with solving inductive reasoning problems that benefit from knowledge of those words (words as tools), or because learning more general words facilitates abstraction (word learning as inductive training).

The current study

We examined whether the types of words in children's vocabulary predict performance on two (putatively nonverbal) inductive reasoning tasks of the kind typically used to assess children's fluid

intelligence. Such tasks require children to look at a sequence of objects, decompose them into constituent dimensions (e.g., shapes, colors, sizes), and extract an abstract pattern in a way that allows children to fill in the missing shape (see Figs. 1 and 2 in Method). The key hypothesis was that children who were reported to produce words with greater generality would perform better at the induction task than children whose vocabularies were of similar size but composed of less general words. We tested this hypothesis using a longitudinal design and by assessing the vocabulary of 2- to 4-year-old children at two timepoints and by seeing whether the makeup of children's vocabulary at Time 1 predicted their performance on induction tasks at Time 2 about 6 months later.

Method

Participants

Participants were 36 children aged 2 to 4 years (mean age at Time 1 = 3.7 years, *SD* = 0.8 years; 21 girls and 15 boys) recruited through the lab's recruitment database and local parent groups.[1] We targeted the 2- to 4-year age range because during this period of development children are rapidly learning new words and this is the youngest age at which it is feasible to test children's inductive reasoning. At Time 1, parents completed our vocabulary checklist at home. At Time 2, parents completed our vocabulary checklist a second time approximately 6 months later and brought their children into the lab to complete the in-lab tasks. The sample was 86% White, 6% Asian, and 8% multiracial. Regarding parents' education, 27% reported having completed a 4-year degree, 24% reported having completed a doctorate, 44% reported having completed a professional degree, 2% reported having completed some college, and 2% chose not to disclose their education history. Parents received a $10 Amazon.com gift card for completing the initial vocabulary survey and an additional $30 in cash for the follow-up lab visit. Children received a book for their participation during the lab visit. One child's data were excluded because the parent indicated that English comprised only 10% of the child's language input.[2]

Materials

We used a combination of surveys and behavioral tasks delivered across two timepoints (see Table 1). Parents initially completed our vocabulary checklist at home. Parents were then contacted approximately 6 months later inquiring whether they were interested in participating in a lab-based portion of the study that occurred 161 to 269 days (*M* = 204 days) after their completion of the initial word checklist.

Vocabulary checklist

We created a vocabulary checklist consisting of 333 words of varying generality covering a range of superordinate, basic-level, and subordinate labels (e.g., *toy* vs. *pinwheel, vegetable* vs. *tomato, number* vs. *five, clean* vs. *vacuum*) to allow us to measure differences in generality across children's reported vocabularies. Below, we briefly describe the methods we used to select the words included on the checklist.

To ensure that there would be ample variance in word knowledge across the 2- to 4-year-olds in our sample, we used the Kuperman norms (Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012) to select words with an estimated age of acquisition (AoA) of 3 to 14 years (mean AoA = 5.60 years). We then excluded any words that were on the MCDI (Fenson et al., 1994; Frank et al., 2017) because 2.5-year-old children tend to produce most of these words. From the remaining words, we used WordNet (G. A. Miller, 1998) to select the 25% most superordinate words as defined by the word's hypernymy—that is, the number of meanings the target word has above it in WordNet's semantic

[1] We initially recruited 169 parents to complete our vocabulary checklist with children in the age range of 2 to 4 years as part of a larger study, but only 41 of these took part in the experiment at Time 2. Of these 41 children, 36 successfully completed all in-lab tasks.

[2] We initially excluded this child for having a PPVT score in the 9th percentile, suggesting a possible language delay (Kelly, 1998). Closer examination revealed the 10% English input (the next lowest value was 60%).

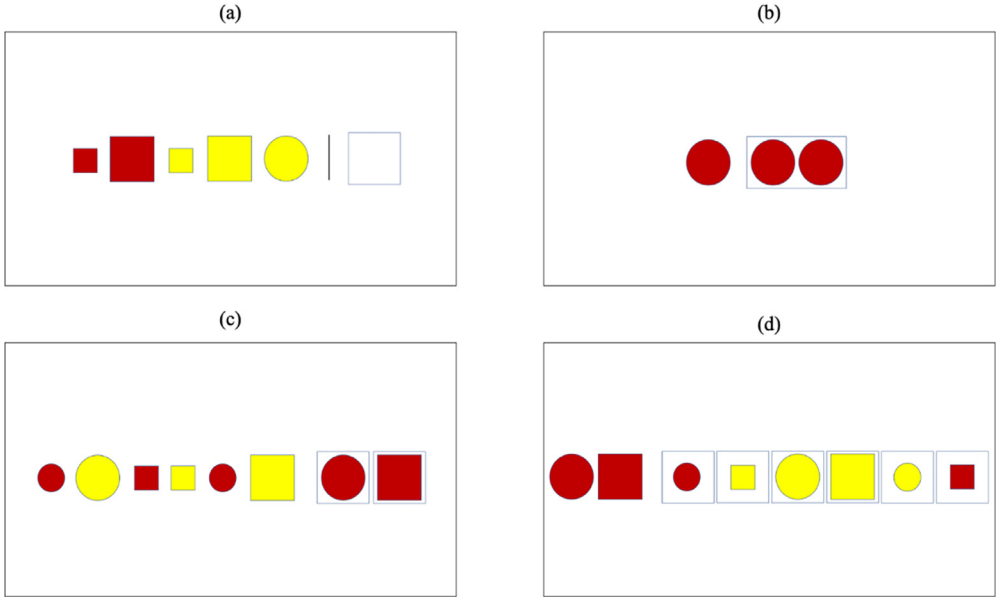

Fig. 1. Examples of verbal/pointing and verbal response-only trials in the Woodcock–Johnson Tests of Cognitive Abilities (WJ) concept formation task (Test 9 of the WJ assessment). (A) Children were asked to "Point to or describe the shape that is the most different and goes in the big box"; the correct answer is "circle." (B–D) Children were asked to explain the rule for a drawing to be inside the box; for example, (B) "two shapes/circles"; (C) "large and red shapes"; (D) "little or yellow shapes." Problems 1 to 5 (A) allowed verbal and/or pointing responses; Problems 6 onward (B–D) required verbal responding.

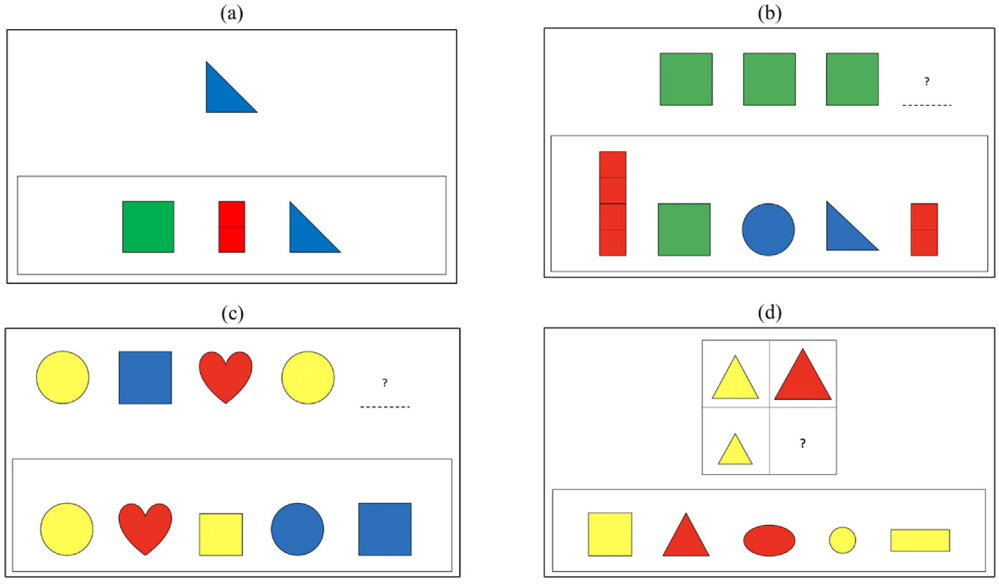

Fig. 2. Examples of perceptual matching and inductive reasoning problems in the Stanford–Binet Intelligence Scales for Early Childhood–Fifth Edition (SB-5) subtest of nonverbal reasoning. Testing began with perceptual match trials of the type shown in Panel A that do not require inductive reasoning. These were followed by simple and complex sequence completion (B and C) and matrices (D).

Table 1
Measures collected from participants at each timepoint.

	Time 1	Time 2
Parents	Vocabulary checklist	Vocabulary checklist
Children	–	In-lab tasks: Word comprehension (PPVT-4);Inductive reasoning (subset of problems from WJ and Early SB-5) ; Color/shape naming

Note. PPVT-4, Peabody Picture Vocabulary Test–Fourth Edition; WJ, Woodcock–Johnson Tests of Cognitive Abilities (Test 9); SB-5, Stanford–Binet Intelligence Scales for Early Childhood–Fifth Edition.

hierarchy. This procedure resulted in a final sample of 75 words (see Table 2 for examples). We refer to these words as "seed words.".

Finally, we used WordNet to find semantically related words with hyponym (i.e., subordinate) relationships to the seed words. We refer to these words as "non-seed words." For example, the seed word *cloth* was included alongside some of its hyponyms *cotton, elastic,* and *flannel.* To ensure that the included non-seed words were the kinds of words some children were likely to know, we excluded words that never occurred in the child-produced speech of the CHILDES corpus (MacWhinney, 2000; Sanchez et al., 2019) and words that occurred very infrequently (log frequency < 2) in the child-directed speech in the same corpus. We next removed any words that overlapped with the previously chosen seed words and words that had disproportionately many (>2 standard deviations above the mean) subordinate words. For the remaining words ($n \sim 1000$), we manually examined the meaning of each word (i.e., the specific meaning of the word that made it subordinate to one of the superordinate words) and excluded words whose subordinate word senses were unlikely to be known by any 2- to 4-year-old. For example, one of the hyponyms of *body* that made it through the above-mentioned filtering criteria was *crown.* But this word was included only because it was a hyponym of *head.* We judged that this sense of *crown* is a sense that no 2- to 4-year-old is likely to know.

The final list included 258 non-seed words, with an average of 4.5 (median = 3) subordinate words for each superordinate word (range = 1–26). Seed words with a relatively large number of subordinate words included *color* (n = 26), *number* (n = 26), *vegetable* (n = 25), and *work* (n = 13).

Table 2 shows some of the words included on our checklist along with their hypernym values, AoA, and log frequency of the words in child-directed speech. The lower the hypernymy score, the higher on the semantic hierarchy the word is. Because verbs reside in much shallower semantic hierarchies

Table 2
Examples of words included in the word checklist and their associated norms.

Word	Type of word	Part of speech[a]	WordNet synset	Word frequency[b]	AoA[c]	Hypernymy (raw)	Hypernymy (scaled by part of speech)
Pinwheel	Non-seed	N	3	2.4	7.6	6	−0.68
Ball	Non-seed	N	1	8.7	2.9	6	−0.68
Toy	Seed	N	1	7.7	3.0	5	−1.20
Number	Seed	N	2	8.1	3.9	4	−1.71
Five	Non-seed	N	1	8.5	4.5	7	−0.16
Color	Seed	N	1	8.8	4.0	5	−1.20
Doorbell	Non-seed		1	4.3	5.1	11	1.90
Color	Seed	V	1	8.8	4.0	1	−0.92
Sweep	Non-seed	V	3	5.4	4.2	2	−0.02
Vacuum	Non-seed	V	1	5.6	6.7	2	−0.02
Clean	seed	V	1	8.2	3.9	1	−0.92
Photograph	Non-seed	V	1	5.1	6.7	4	1.79

[a] Part of speech of the relevant WordNet synset. Only nouns (N) and verbs (V) are organized hierarchically in WordNet.
[b] Log-transformed frequency of the word in child-directed speech. These are word form based and do not distinguish between parts of speech.
[c] Adult-produced estimates of age of acquisition (AoA; Kuperman et al., 2012).

than nouns, their maximum hypernymy values tend to be smaller than those for nouns. In our main analyses, therefore, we scaled hypernymy by part of speech such that 0 corresponds to a noun or verb of average hypernymy and − 1 corresponds to a noun or verb that is 1 standard deviation more general (higher on the semantic hierarchy) than the average noun or verb. These scaled values are shown in Table 2.

Properties of the words on the checklist. A full list of the words and associated norms is available in the online repository on the Open Science Framework (https://osf.io/chz7w). A comparison of the seed and non-seed words on some key lexical characteristics is shown in Table 3.

On average, seed words had—as intended—significantly fewer hypernyms, were more polysemous as measured by the number of WordNet synsets, had higher frequencies in child-directed speech, had a lower AoA, and were slightly more abstract. Given these relationships, one might wonder whether hypernymy can be reduced to these more familiar lexical measures. In a multiple regression predicting hypernymy (scaled by part of speech) from concreteness, frequency, AoA, and (logged) number of synsets, we found that, taken together, these predictors account for only 8% of the variance, with polysemy (log-transformed number of synsets) as the only reliable predictor (b = −.23, 95% confidence interval (CI) = [−.38, −.10], t = 3.30, p <.001). Removing the number of synsets reduces R^2 to less than 5% and unmasks a significant (although small) effect of frequency; controlling for concreteness and AoA, higher hypernymy (i.e., greater specificity) is associated with lower frequency (b = −.10, 95% CI = [−.19, −.02], t = − 2.40, p =.02). In short, it is not the case that hypernymy can be reduced to these other predictors.

To further validate our measure of hypernymy as a psychological construct, we verified that it correlates with adult judgments of word generality—that is, how general versus specific a word meaning is (for more details, see Lewis, Colunga, & Lupyan, 2021).[3] A remaining concern is that hypernymy as quantified here is not a very accurate measure of semantic generality as represented by children because the WordNet hierarchies on which it is based are characterized by expert knowledge, bearing only a passing relationship to children's (and even many adults') semantic hierarchies. This is a valid concern, also affecting many other studies that rely on adult semantic features (Hills, Maouene, Maouene, Sheya, & Smith, 2009) or word associations (Steyvers & Tenenbaum, 2005) to stand in for children's concepts. These imperfect measures are, nevertheless, useful to the extent that they allow us to measure and predict children's development and are not readily quantifiable for children's conceptual and lexical knowledge.

Administration of the word checklist. Parents completed the word checklist at home in a web browser at Time 1 and either at home or in the lab at Time 2. They separately indicated for each word whether their children understood it and whether they produced it. Words were presented using the same categories and category order as the MCDI (e.g., small household objects, action words; Fenson et al., 1994; Frank et al., 2017). Although parents always saw the categories of words in the same order, the order of individual words within each category was randomized.

In-lab behavioral tasks

PPVT–Fourth Edition. As the first in-lab task, we administered the PPVT–Fourth Edition (PPVT-4) as a standardized measure of word comprehension (Dunn & Dunn, 2007). The test was administered via printed booklet.

[3] Our word generality ratings are included in the online repository at the Open Science Framework (https://osf.io/6b4hm). As in earlier work (Lewis et al., 2021), words rated by adults as having more general meanings had significantly fewer hypernyms, but this measure was also—and to a much stronger extent—correlated with the word's number of senses (WordNet synsets) and hyponymy (i.e., number of synsets that are "below" the target word). More superordinate (general) words had more synsets and hyponyms, suggesting that when asked about a word's generality, people conflate hypernymy, polysemy, and the word's semantic density. As it turns out, the hypernymy of words a child knows is related to later outcomes as we measure them, whereas polysemy and semantic density are not. For this reason, our analyses use WordNet hypernymy rather than subjective ratings of generality.

Table 3
Characteristics of seed versus non-seed words making up the vocabulary checklist.

Characteristic	Word type		
	Seed	Non-seed	Comparison
Number of words	75	258	–
Mean number of hypernyms	3.36	6.58	$t = -10.22$, $p < .001$
Hypernymy scaled by part of speech	−1.05	0.31	$t = -13.30$, $p < .001$
Mean number of synsets	7.81	4.64	$t = 3.17$, $p = .002$
Mean log frequency[a]	6.92	5.74	$t = 5.74$, $p < .001$
Mean AoA	4.72	5.87	$t = -5.61$, $p < .001$
Mean concreteness	3.81	4.04	$t = -2.20$, $p = .03$

Note. Seed words had lower hypernymy; that is, they had more general meanings. Unsurprisingly, seed words had lower concreteness and a larger number of meanings (synsets). Somewhat unexpectedly, seed words also had significantly lower age of acquisitions (AoAs) and greater frequency in child-directed speech.

[a] Frequency is log-transformed counts of the words in child-directed U.S./U.K. English speech in the CHILDES corpus.

Inductive reasoning. We administered two reasoning tests: the concept formation test (Subtest 9) from the Woodcock–Johnson Tests of Cognitive Abilities (WJ; Woodcock & Johnson, 1989; see Fig. 1 for examples) and the sequence and matrix problems for nonverbal reasoning from the Stanford–Binet Intelligence Scales for Early Childhood–Fifth Edition (SB-5; Roid & Pomplun, 2012; see Fig. 2 for examples). Inductive reasoning has long been hypothesized to be at the core of fluid reasoning (e.g., Carroll, 1993). Both tasks aim to assess children's ability to observe a phenomenon such as a series of shapes and discover the underlying rule or principle that is responsible for giving rise to it.

WJ (concept formation subtest). Problems 1 to 5 allowed verbal or pointing responses (i.e., children could point to items). Problems 6 onward required explicitly verbal responses and involved progressively more complex verbal solutions (e.g., Problems 21–29 required children to respond using logical operators *and* and *or*). Of the two inductive reasoning tasks, therefore, the WJ is a more verbal test of inductive reasoning (see "Procedure" section for more details on how the different types of problems were administered).

SB-5 (subset of nonverbal reasoning). Next, we used nine inductive reasoning problems from the nonverbal reasoning section of the Early SB-5 intelligence test (Roid & Pomplun, 2012), preceded by three simple perceptual matching problems to familiarize children with the task (we did not administer the entire SB-5 reasoning section because it would have been too time-consuming). The shapes were circles, squares, triangles, rectangles, and hearts; the colors were blue, yellow, red, green, and pink. Sequence trials presented children with a sequence of shapes of different colors and asked them to select which shape best completed the sequence. Matrix trials presented children with a 2×2 grid filled with three shapes and asked them to fill in the missing element (i.e., the fourth shape).

Color and shape naming. Lastly, we assessed children's expressive knowledge of the colors and shapes used in the two reasoning tests. The colors (i.e., blue, yellow, red, green, and pink) and shapes (i.e., square, circle, triangle—equilateral and right-angled, rectangle, oval, and heart) were those used in the WJ and the SB-5. The squares appeared in 5 different trials (i.e., a single square and also two, four, six, and eight squares stacked on top of one another). Triangles appeared in 2 different trials (i.e., equilateral and right-angled triangles). Rectangles appeared in 2 different trials (i.e., vertical and horizontal configurations). The rest of the stimuli appeared once. Hence, there were 17 trials in total.

Procedure

First session (Time 1)

Parents completed a demographic questionnaire and the vocabulary checklist online to measure the characteristics of known words at Time 1.

Second session (Time 2)

Approximately 6 months after Time 1, parents completed the same vocabulary checklist as in Time 1— this time to measure the characteristics of known words at Time 2. Parents had the option to complete the vocabulary checklist either during the lab session or at home prior to coming to the lab. During the lab session, children first completed the PPVT-4, followed by the inductive reasoning tasks (i.e., WJ and subset of SB-5) and then the color and shape naming task. Children were tested individually in a quiet room in the lab with their parents either in the waiting room or seated out of sight in the testing room.

PPVT-4. The PPVT-4 was administered as a control for overall vocabulary. It was administered following the standard instructions and was used to control for overall language knowledge. Children were shown a series of images in 2 $\times$ 2 grids and were asked to point to a target object (e.g., "Point to the carrot"). Testing concluded when children incorrectly responded to eight or more prompts within a testing block. Responses were scored and normed by age per the instruction manual, yielding a standard score and a percentile for each child.

WJ (concept formation subtest). The WJ was administered as a test of induction ability, following the standard instructions. Children were shown a series of images and were asked to point to or say which object belonged in the big box or why an object belonged in the big box, indicating that it differed from the other objects shown within each problem. Whereas Problems 1 to 5 allowed verbal or pointing responses (i.e., children could point to items), Problem 6 onward required explicitly verbal responses and involved progressively more complex verbal solutions (e.g., Problems 21–29 required children to respond using logical operators *and* and *or,* whereas Problems 30–40 included a mixture of problem types). Standard instructions included cutting off administration based on performance. For example, if children responded correctly on fewer than 3 problems from the first set (Problems 1–5), they did not advance to the next set. In our sample, children received a minimum of 5 problems and a maximum of 40 problems.

SB-5 (subset of nonverbal reasoning). The SB-5 was administered as a second test of induction ability. The experimenter told children that they were going to play a game with shapes and colors on an iPad. The experimenter introduced the task by showing children the first trial and saying, "Let's play the 'find it' game—I'm going to find one just like this," while pointing to the target shape in the center of the screen. The experimenter then gestured to the answer options at the bottom of the screen, pointed to the correct response, and said, "See? This one is just like the other one." Next, the experimenter pointed to the target shape in the center of the screen and asked children to point to the answer, saying, "Now you do it. Point to the one that looks just like this." On subsequent shape matching trials (Trials 2 and 3), the experimenter simply said, "Point to the one like this." In the remaining trials, the experimenter first drew children's attention to the row or grid of shapes at the center of the screen, then pointed to the question mark and said, "Something is missing here." The experimenter then gestured along the answer choices at the bottom of the screen and told children, "Point to the one that should go here." During the entire task, the experimenter did not name any shapes or colors of the stimuli. Children's responses were self-paced and recorded by the experimenter. All children received the problems in the same order—beginning with simple sequences, proceeding to more complex sequences and matrices, and ending with the most complex sequences.

Color and shape naming. The color and shape naming task was administered to ensure that children did not need to be excluded from the inductive reasoning tasks based on their inability to verbally name shapes and colors. The experimenter first told children that they were going to play a game about colors. For each trial in the color task, the experimenter asked children "What color is this?" and recorded the response. The experimenter then repeated the procedure for the shape task, asking "What shape is this?" and recording the response as correct or incorrect for both tasks. Trials were always presented in the same order.

Results

Descriptive statistics

PPVT-4

To confirm that our participants showed typical language development, we first examined their PPVT scores. Children's PPVT performance was largely above published norms (mean percentile = 83.34, *SD* = 13.04; mean standard score = 118.06, *SD* = 9.63).

Vocabulary checklist

At Time 1, children were reported to say an average of 227 of 333 words (*SD* = 54, min = 110, max = 316) on the checklist. At Time 2, (*M* = 6.8 months later, *SD* = 1.23), children were reported to say significantly more words (*M* = 257, *SD* = 43, min = 176, max = 318), $t(34) = 7.20$, $p < .001$. The same pattern was observed for understanding. At Time 1, children were reported to understand an average of 224 of 333 words (*SD* = 73, min = 26, max = 309)[4] on the checklist. At Time 2, children were reported to understand significantly more words (*M* = 257, *SD* = 51, min = 97, max = 314), $t(34) = 2.77$, $p = .009$, than at Time 1. See online supplementary material for scatterplots showing the numbers of words reported as "says" and "understands" across Time 1 and Time 2.

Based on each word on the checklist that children were reported to say, we calculated the mean word frequency (based on child-directed speech in the CHILDES corpus), AoA, and hypernymy of each child's vocabulary at Time 1 and Time 2. When calculating mean hypernymy, we first obtained the hypernym value for each word that the parent checked off. Because the number of hypernyms a word has differs by the word's part of speech (i.e., nouns on average have much lower hypernymy scores than verbs),[5] we normalized hypernym values by part of speech (i.e., noun or verb) prior to averaging. See supplementary material for raw hypernymy values for all words.

Mean hypernymy of a given child's vocabulary was simply the average of all the words from the checklist that the child was reported to produce. Greater mean hypernymy—as we defined it—is distinct from the *number* of words the child knows; a child whose vocabulary includes only very general words (low hypernymy) would have a lower hypernymy value than a child whose vocabulary includes more words but only very specific ones (high hypernymy value).

Inductive reasoning performance on WJ (concept formation subtest)

Because the test is self-terminating, different children saw different numbers of problems. On average, children saw 12 problems (median = 11) before the test terminated due to multiple incorrect responses. Children demonstrated substantial variability in performance, responding to an average of 5.20 problems correctly (range = 0–27, *SD* = 5.84). This suggests that the task was quite difficult for them. In retrospect, this was not surprising considering that the children in our sample were of the lowest age range for which these problems are designed. Because raw score performance was heavily skewed toward low numbers (skewness = 2.13), we used the ranks of the raw scores as the outcome variable.[6]

Inductive reasoning performance on SB-5 (subset of nonverbal reasoning)

Our outcome measure is the number of questions answered correctly out of the 9 induction trials completed by each child. Children demonstrated substantial variability in performance (*M* = 5.57 of 9 trials correct, *SD* = 2.46, range = 1–9). As we describe below, about 40% of the variance is explained by age.

[4] We needed to use existing "says" and "understands" scores from 35 children to impute the total number of words understood at Time 1 and Time 2 for 1 child due to a missing data point in a parent's reporting of how many words the child understood.

[5] The large difference in hypernymy between nouns and verbs is a consequence of how WordNet is organized. Nouns tend to have much deeper hierarchies than verbs, which correspond to a larger maximum hypernymy value.

[6] Qualitatively similar results are obtained if we use the grade-equivalent scores that are a nonlinear transform of the raw scores; that is, scores below 7 are below kindergarten (coded as 0), whereas the highest score in our group—27—is equivalent to grade level 5.4.

Color and shape naming

Children performed well on both the shape naming task (M = 5.97 of 8 trials correct, SD = 2.09, range = 1–8) and color naming task (M = 4.69 of 5 trials correct, SD = 0.47, range = 4–5), demonstrating that they were familiar with the shapes/colors used in the inductive reasoning tasks. No children needed to be excluded based on their color and shape naming performance.

Socioeconomic status

There is good reason to think that performance on inductive reasoning tests is related to parental SES (Ardila, Rosselli, Matute, & Guajardo, 2005; Bradley & Corwyn, 2002; Brooks-Gunn & Duncan, 1997; Hart, Petrill, Deckard, & Thompson, 2007). If so, it is important to know whether the correlation between hypernymy and inductive reasoning remains when parental SES is taken into account. For example, perhaps higher-SES parents are more likely to use abstract language that helps children to learn it and such children also happen to perform better on inductive reasoning tasks for reasons perhaps entirely unrelated to their vocabulary or linguistic environment. Controlling for SES allows one to consider this possibility. We operationalized SES as a summed standardized score of parental education and income.

Relationship among predictors

Fig. 3 shows correlations among the by-child predictors. Unsurprisingly, some measures of children's vocabulary (e.g., mean frequency, mean AoA) are very strongly correlated. This is important to know for avoiding collinearity in the analyses presented below.

Relationships between language and inductive reasoning

We present separate analyses for the two induction tasks we administered because although performance on them was moderately correlated, they relate to children's vocabulary size and composition in rather different and theoretically interesting ways. In the supplementary materials, we have included plots showing the relationship of vocabulary hypernymy to individual performance on both the SB-5 and WJ induction tasks.

SB-5 (subset of nonverbal reasoning). We began with a baseline logistic regression model predicting performance (i.e., proportion of 9 problems solved) from the child's age at Time 2 (see Table 4, Model 1). We next added the control variables SES and the total number of words produced by the child on our checklist at Time 1 (Table 4, Model 2). We then added to the model the child's PPVT standard score at Time 2 (see Table 4, Model 3). Across Models 1 to 3, only age and SES were significant predictors of performance, accounting for about 46% of the variance. We next added the mean hypernymy of the child's vocabulary at Time 1 (recall that a lower mean hypernymy score corresponds to a child knowing more superordinate words on average). This score was significantly related to inductive reasoning, accounting for an additional 8% of the variance; children whose vocabulary comprised more superordinate words performed better on an inductive reasoning task administered more than 6 months later (see Table 4, Model 4). We performed the same analyses looking at whether the hypernymy of only known nouns or known verbs predicts performance on the SB-5. The effects of hypernymy were driven more by knowledge of more general verbs than by knowledge of more general nouns. However, there was not a significant interaction by part of speech (CI for noun-only analysis = [−0.75, 0.40]; CI for verb-only analysis = [−0.69, 0.41]), suggesting that both nouns and verbs contribute to hypernymy and its association with SB-5 performance.

One possibility may be that knowing *vacuum* but not *clean* is symptomatic of an atypical language learning trajectory (e.g., see Beckage, Smith, & Hills, 2011, for a related analysis). If so, any differences on inductive reasoning observed for such children may indicate poorer learning abilities in general. Two analyses speak against this interpretation. First, although hypernymy was correlated with the total number of words children were reported to say at Time 1, $r(33)$ =.39, p =.02, the positive correlation indicates that children with a larger productive vocabulary knew words with slightly higher mean hypernymy (i.e., their vocabulary was skewed toward more specific words). Second, the vocabulary of poorer word learners should be skewed toward words with earlier AoAs and/or more frequent words. Yet hypernymy continued to be a significant predictor of SB-5 performance when we also

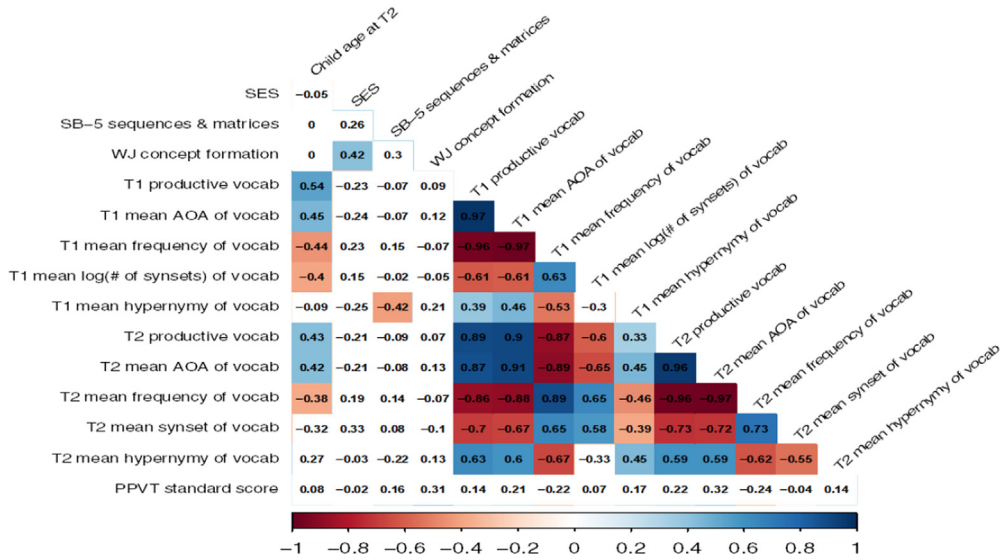

Fig. 3. Pearson correlations among age, task performance, and vocabulary characteristics at Time 1 and Time 2. Correlations with the two inductive reasoning tasks (Stanford–Binet Intelligence Scales for Early Childhood–Fifth Edition [SB-5] and Woodcock–Johnson Tests of Cognitive Abilities Woodcock–Johnson Tests of Cognitive Abilities [WJ]) partial out age at the time of testing. Statistically significant correlations are indicated by colored squares. SES, socioeconomic status; T1, Time 1; T2, Time 2; AOA, age of acquisition; vocab, vocabulary; PPVT, Peabody Picture Vocabulary Test. (For interpretation of the reference to color in this figure legend, the reader is referred to the Web version of this article.).

Table 4
Relationship between inductive reasoning assessed by SB-5 and age at Time 2, SES, total vocabulary size at Time 1, PPVT standard score, mean hypernymy of vocabulary at Time 1, mean AoA of vocabulary at Time 1, and mean frequency of vocabulary at Time 1.

Model	Predictor	*b*	Standard 95% CI	*z*	*p*	d^2
Model 1	Time 2 age	0.35	−0.11, 0.84	6.44	<.001	.42
Model 2	Time 2 age	0.35	−0.19, 0.90	5.75	<.001	.46
	SES	0.08	−0.36, 0.59	2.04	.041	
	Time 1 total vocabulary	0.00	−0.54, 0.56	−0.23	.815	
Model 3	Time 2 age	0.34	−0.19, 0.90	5.77	<.001	.48
	SES	0.08	−0.36, 0.59	2.02	.043	
	Time 1 total vocabulary	−0.01	−0.56, 0.56	−0.36	.718	
	Time 2 PPVT standard score	0.05	−0.42, 0.52	1.42	.156	
Model 4	Time 2 age	0.27	−0.28, 0.86	4.37	<.001	.56
	SES	0.05	−0.42, 0.57	1.41	.160	
	Time 1 total vocabulary	0.12	−0.53, 0.76	0.97	.330	
	Time 2 PPVT standard score	0.08	−0.40, 0.56	1.74	.082	
	Time 1 mean hypernymy	−0.21	−0.75, 0.35	−2.98	.003	
Model 5	Time 2 age	0.28	−0.24, 0.84	4.67	<.001	.56
	SES	0.05	−0.42, 0.56	1.42	.156	
	Time 2 PPVT standard score	0.07	−0.41, 0.56	1.62	.105	
	Time 1 mean hypernymy	−0.22	−0.79, 0.35	−3.03	.002	
	Time 1 mean AoA	0.13	−0.51, 0.75	1.11	.268	
Model 6	Time 2 age	0.29	−0.24, 0.86	4.82	<.001	.56
	SES	0.04	−0.42, 0.56	1.36	.173	
	Time 2 PPVT standard score	0.07	−0.41, 0.56	1.69	.092	
	Time 1 mean hypernymy	−0.21	−0.80, 0.40	−2.74	.006	
	Time 1 mean frequency	−0.10	−0.77, 0.57	−0.57	.571	

Note. SB-5, Stanford–Binet Intelligence Scales for Early Childhood–Fifth Edition; SES, socioeconomic status; PPVT, Peabody Picture Vocabulary Test; AoA, age of acquisition; CI, confidence interval.

controlled for the mean AoA of the words reported to be known at Time 1 (see Table 4, Model 5; alternatively, see the mean log frequency of the produced words in Table 4, Model 6). Neither mean AoA ($t < 2$) nor mean word frequency ($t < 1$) of the known words was predictive of SB-5 performance,[7] as would be predicted if the link between hypernymy and reasoning performance simply reflected differences in word learning abilities.

Whereas SES was predictive of performance in Models 2 and 3 (albeit a relatively weak association, $ts = 2.02–2.04$), SES was no longer predictive of performance once hypernymy of vocabulary at Time 1 was included in the models.

Performance on the SB-5 problems was not predicted by language measures assessed at Time 2 (i.e., on the same day as the reasoning problems were administered); that is, SB-5 performance was not predicted by the total number of words checked on our list ($ts < 1$) or by the hypernymy measure at Time 2 ($ts < 2$). We speculate about what this means in the Discussion. Performance at Time 2 was predicted by age ($ts > 6$) and SES ($ts > 2$) (see supplementary material for Time 2 analyses).

WJ subtest (concept formation). The base model used children's age at Time 2 as a predictor (Table 5, Model 1) and progressively added predictors to the model in the same order as the SB-5 analysis. As evident from Models 1 to 6 in Table 5, age was a significant predictor of performance ($ts > 5$), which accounted for about 53% of the variance. In addition, Models 2 to 6 show that SES was a significant predictor of performance on the WJ ($ts > 2.5$), accounting for an additional 8% of variance. Inductive performance on the WJ task was not predicted by total vocabulary, or the PPVT standard score, or by the mean hypernymy, AoA, or frequency of the vocabulary at Time 1 or Time 2 ($ps > .05$) (see supplementary material for Time 2 analyses).

Recall that the WJ concept formation subtest contains two types of problems: the initial (easier) problems allowing pointing or verbal responses (Problems 1–5) and later problems allowing only verbal responses (Problems 6 onward). We examined whether vocabulary hypernymy predicted performance differently for problems that allowed pointing/verbal responses or verbal responses only; neither was significantly associated with vocabulary hypernymy ($ps > .05$). We also examined the correlation between performance on WJ Problems 1 to 5 (i.e., which allowed pointing or verbal responses) and the SB-5 and the correlation between performance on WJ Problems 6 onward (i.e., verbal responses only) and the SB-5. There was not a significant correlation between performance on the SB-5 and either WJ Problems 1 to 5 ($r = .12$, $p = .48$) or WJ Problems 6 onward ($r = .24$, $p = .17$). See supplementary material for additional analyses.

Discussion

Our goal was to investigate whether children's knowledge of certain types of words was related to their inductive reasoning. We hypothesized that having a vocabulary consisting of more superordinate words would be associated with better performance in inductive reasoning. We observed the predicted association consistently for one of our induction tasks (see Table 4). The association held after controlling for children's overall vocabulary size and parental SES. We found an association between vocabulary hypernymy and inductive reasoning as tested by a more nonverbal test (the sequence and matrix problems from the SB-5; see Table 4), although not for a more linguistically loaded reasoning test (the concept formation section of the WJ; see Table 5) that required children to verbally articulate the rules governing patterns of shapes and colors. Whether this difference between the two tests is meaningful, reflecting perhaps the more verbal nature of the WJ problems, requires further testing.

Why might learning and using more superordinate words be associated with better performance on some inductive reasoning tasks? One possibility, which we referred to earlier as the *common cause* hypothesis, is that producing more general words is a sign of children's more developed abstraction ability—the same ability that leads to better performance on nonverbal inductive reasoning tasks.

[7] Models that included mean AoA and frequency of vocabulary at Time 1 did not include total vocabulary at Time 1 due to high correlations between these factors (i.e., mean AoA and total vocabulary correlate at.97; mean frequency and total vocabulary correlate at −.96).

Table 5
Relationship between inductive reasoning assessed by WJ and age at Time 2, SES, total vocabulary size at Time 1, PPVT standard score, mean hypernymy of vocabulary at Time 1, mean AoA of vocabulary at Time 1, and mean frequency of vocabulary at Time 1.

Model	Predictor	*b*	Standard 95% CI	*t*	*p*	Adjusted R^2
Model 1	Time 2 age	0.73	0.49, 0.97	6.22	<.001	.53
Model 2	Time 2 age	0.66	0.39, 0.92	5.12	<.001	.61
	SES	0.32	0.09, 0.55	2.88	.007	
	Time 1 total vocabulary	0.18	−0.09, 0.44	1.34	.189	
Model 3	Time 2 age	0.65	0.40, 0.90	5.32	<.001	.64
	SES	0.32	0.10, 0.53	2.97	.006	
	Time 1 total vocabulary	0.15	−0.11, 0.41	1.16	.254	
	Time 2 PPVT standard	0.20	−0.01, 0.41	1.91	.065	
Model 4	Time 2 age	0.72	0.46, 0.99	5.55	<.001	.65
	SES	0.34	0.12, 0.56	3.21	.003	
	Time 1 total vocabulary	0.05	−0.24, 0.34	0.37	.716	
	Time 2 PPVT standard	0.18	−0.03, 0.39	1.72	.096	
	Time 1 mean hypernymy	0.17	−0.08, 0.42	1.42	.165	
Model 5	Time 2 age	0.73	0.48, 0.98	5.94	<.001	.65
	SES	0.34	0.12, 0.56	3.21	.003	
	Time 2 PPVT standard	0.18	−0.04, 0.39	1.68	.103	
	Time 1 mean hypernymy	0.17	−0.08, 0.43	1.37	.181	
	Time 1 mean AoA	0.05	−0.23, 0.34	0.36	.719	
Model 6	Time 2 age	0.78	0.52, 1.03	6.24	<.001	.65
	SES	0.33	0.12, 0.55	3.14	.004	
	Time 2 PPVT standard	0.18	−0.03, 0.40	1.77	.088	
	Time 1 mean hypernymy	0.22	−0.05, 0.49	1.67	.105	
	Time 1 mean frequency	0.05	−0.25, 0.35	0.36	.723	

Note. WJ, Woodcock–Johnson Tests of Cognitive Abilities (Test 9); SES, socioeconomic status; PPVT, Peabody Picture Vocabulary Test; AoA, age of acquisition; CI, confidence interval.

Although we cannot fully rule out this possibility, this account predicts a relationship between vocabulary hypernymy and overall vocabulary size—children who are better at abstracting should be better at learning words in general—which we do not observe here. Further evidence against the common cause explanation is a recent result from Lewis et al. (2021), who found that a positive relationship between knowledge of superordinate-type words and a subsequent increase in the rate of word learning remained even after controlling for fluid reasoning scores.

If the link between producing more general words and solving inductive reasoning problems is causal, this may be because knowing specific superordinate words like *color* may help children to reason about problems involving color sequences (words as tools). Given that only a small number of words included on the checklist appear to be related to the specific inductive reasoning problems that the children needed to solve (e.g., the words *color* and *shape*), it is unclear that it is knowledge of *specific* words that is making the difference here. More plausibly, using more general words may provide children with increased practice—a form of cognitive training—in abstracting at a higher level—treating highly heterogeneous objects and actions as similar (Christie & Gentner, 2010; Gentner & Namy, 1999). This may promote structural alignment and help children to formulate higher-order relational hypotheses (e.g., "these are all . . .," "this one's not like the others because . . .") of the sort useful for solving inductive reasoning problems like those tested here (see Figs. 1 and 2). The learning and use of superordinate words may share a mechanism responsible for a recent finding by Simms and Richland (2019), who showed that eliciting relational language from preschoolers (which often involved the use of superordinate words like *do* and *make*) improved their performance in a picture-based analogical reasoning task. Our result also bears some similarity to work by Frausel et al. (Frausel, Richland, Levine, & Goldin-Meadow, 2021; Frausel et al., 2020), who studied what they call HOTT language (higher-order thinking talk), defined as talk that includes "reference to an inference or explanation, a comparison, an abstraction/generalization, or a hierarchy/taxonomic relationship" (Frausel et al., 2020, p. 2). Greater use of HOTT by children aged 14 to 58 months predicted better analogical/inductive reasoning when the children were tested at 9 and 11 years (Frausel

et al., 2020). Certainly, HOTT does not require the use of superordinates, but we suspect that many of the utterances Frausel et al. coded as instances of HOTT make use of relatively superordinate words. Earlier learning of such words may be associated with greater use of HOTT.

Another possibility is that children who know more superordinate words happen to be exposed to an environment that fosters the learning of such words, and it is this environment—rather than their vocabulary knowledge—that is responsible for better inductive reasoning performance. Children in such environments may be more often tasked with categorizing at higher and more relational levels and encouraged to explain how different objects and events are related—that is, tasks that create a greater communicative need for more superordinate words, which in turn helps children to learn them. These may be the same types of environments that promote higher-order thinking talk (Frausel et al., 2021). It remains an open question whether the benefits of such environments for inductive reasoning accrue independent of the use of superordinate terms or whether superordinate terms comprise an important (and perhaps sufficient) proximate mechanism.

Limitations and remaining questions

A clear limitation of our study is its small sample size. This limitation concerns the possibly low power associated with our results and the possibility that the reported results over-estimate the true effect size and even mis-estimate its direction (Gelman & Carlin, 2014). We did not conduct a priori power analysis; a post hoc power analysis using the effect of Time 1 hypernymy on SB-5 performance as the key finding yields a power estimate of.82 for Model 4 and.74 for Model 6. The corresponding Type M errors that capture the over-estimation of the magnitude of the observed effect are 1.11 and 1.16; that is, our design is likely to over-estimate the true effect by 11% to 16%. The corresponding Type S errors—the likelihoods that the true effect is in the opposite direction of what is observed here—is less than.0001 (calculations done using the *retrodesign* R package; Timm, Gelman, & Carlin, 2019).[8]

Another limitation concerns the relative homogeneity of the participants, who were largely White middle- and upper-SES families (86%). On the one hand, finding individual differences in a sample with such restricted range suggests that the effect may be larger in a population with wider variance in hypernymy and inductive reasoning. On the other hand, collecting data from a more diverse sample would help to determine whether effects of hypernymy on inductive reasoning generalize broadly. Another limitation is the short interval between Time 1 and Time 2; the use of just two timepoints limits our ability to examine cross-lagged correlations that can help to support causal inferences. Relatedly, the absence of a reasoning baseline test at Time 1 prevents us from comparing the relationship between Time 1 reasoning to Time 2 hypernymy and Time 1 hypernymy to Time 2 reasoning. Finding that hypernymy predicts later reasoning more than the reverse, and that hypernymy predicts subsequent reasoning after controlling for Time 1 inductive reasoning ability, would further help to clarify causal links between these measures.

If vocabulary hypernymy—greater knowledge of more superordinate words—is causally linked to inductive reasoning, why was inductive reasoning predicted by vocabulary hypernymy across a timespan of more than 6 months but not by a contemporaneous measure of vocabulary hypernymy (i.e., hypernymy measured at Time 2)? We do not have a satisfying answer. One possibility is that the hypernymy measure at Time 2 is less sensitive than the measure at Time 1 because children become more similar to one another on the word checklist (i.e., a saturation effect). Arguing against this possibility is our finding that although children do become somewhat more similar in terms of their productive vocabulary size (Time 1 *SD* = 54, Time 2 *SD* = 43), there was little change in the variability of mean hypernymy (Time 1 *SD* =.12, Time 2 *SD* =.13). Another possibility is that word knowledge as indexed through checklists may, at least initially, indicate the use of a word in a highly restricted context. This range of contexts will expand during the months subsequent to the checklist administration, and it is this change (which we do not currently have a way to measure) that is causally linked to

[8] If we are over-estimating the true effect by 15%, then our achieved power falls to.69, the Type S error is still less than.0001, and the Type M error increases to 1.63. If we are over-estimating the effect by a factor of 2, then power falls to.29, Type S error increases to.002, and Type M error increases to 2.77.

inductive reasoning performance. If true, then hypernymy measured at Time 2 may predict reasoning performance at a future time—for example, Time 3—even after controlling for reasoning performance at Time 2.

We found a link between vocabulary hypernymy and inductive reasoning when using the matrices and sequences questions of the SB-5 but not when using the WJ concept formation test—a more linguistically loaded reasoning test that required children to articulate the induced rules. We suspect that this discrepancy is due to the restricted range in the performance on the WJ. With the exception of a few children, performance was quite low. Although the problems are validated for use in 2- to 4-year-olds, children at this age range are expected to perform essentially at floor, which—except for a few children—is what we found.

Conclusions

Our results show that knowledge of more general words is positively associated with inductive reasoning in 2- to 4-year-old children, at least when using the sorts of nonverbal problems shown in Fig. 2B to 2D. The benefit of knowing more superordinate words could not be attributed to these children simply having larger vocabularies or knowledge of rarer words. Our findings stress the importance in measuring not just *how many words* children are reported to know but also *what sorts of words* they know when measuring the influence of vocabulary on several important outcomes in later life. Although our experimental design does not allow us to draw causal conclusions, the results are consistent with the possibility that early knowledge of seed words can facilitate performance on some nonverbal inductive reasoning tasks. The most direct way to distinguish between the possibilities outlined above is to intervene in children's knowledge of superordinate words through, for example, directed instruction of either superordinate or semantically related, more specific words (e.g., Neuman, Newman, & Dwyer, 2011) and then testing subsequent inductive reasoning performance. Such a training study would allow for isolating the effect of learning superordinate words while keeping constant children's broader environment.

Acknowledgments

This study received funding from a National Institutes of Health R21 grant (HD092867) and a UW-2020 award to G.L. and H.A.V. We are grateful for the contribution of Alexis Hosch and to all children and caregivers who took part in our research.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jecp.2022.105449.

References

Ardila, A., Rosselli, M., Matute, E., & Guajardo, S. (2005). The influence of the parents' educational level on the development of executive functions. *Developmental Neuropsychology, 28*(1), 539–560.

Balaban, M. T., & Waxman, S. R. (1997). Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology, 64*, 3–26.

Beckage, N. M., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLoS One, 6*(5) e19348.

Bradley, R. H., & Corwyn, R. F. (2002). Socioeconomic status and child development. *Annual Review of Psychology, 53*, 371–399.

Brooks-Gunn, J., & Duncan, G. J. (1997). The effects of poverty on children. *The Future of Children, 7*(2), 55–71.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies*. Cambridge, UK: Cambridge University Press.

Casasola, M. (2005a). Can language do the driving? The effect of linguistic input on infants' categorization of support spatial relations. *Developmental Psychology, 41*, 183–192.

Casasola, M. (2005b). When less is more: How infants learn to form an abstract categorical representation of support. *Child Development, 76*, 279–290.

Christie, S., & Gentner, D. (2010). Where hypotheses come from: Learning new relations by structural alignment. *Journal of Cognition and Development, 11*, 356–373.

Clark, A., & Karmiloff-Smith, A. (1993). The cognizer's innards: A psychological and philosophical perspective on the development of thought. *Mind & Language, 8*, 487–519.
Davidoff, J., & Roberson, D. (2004). Preserved thematic and impaired taxonomic categorisation: A case study. *Language and Cognitive Processes, 19*, 137–174.
Deng, W., & Sloutsky, V. M. (2013). The role of linguistic labels in inductive generalization. *Journal of Experimental Child Psychology, 114*, 432–455.
Duff, F. J., Reen, G., Plunkett, K., & Nation, K. (2015). Do infant vocabulary skills predict school-age language and literacy outcomes? *Journal of Child Psychology and Psychiatry and Allied Disciplines, 56*, 848–856.
Dunn, L. M., & Dunn, D. M. (2007). *PPVT-4: Peabody Picture Vocabulary Test*. San Antonio, TX: Pearson Assessments.
Edmiston, P., & Lupyan, G. (2015). What makes words special? Words as unmotivated cues. *Cognition, 143*, 93–100.
Edwards, L., Figueras, B., Mellanby, J., & Langdon, D. (2011). Verbal and spatial analogical reasoning in deaf and hearing children: The role of grammar and vocabulary. *Journal of Deaf Studies and Deaf Education, 16*, 189–197.
Fenson, L., Cameron, M. S., & Kennedy, M. (1988). Role of perceptual and conceptual similarity in category matching at age two years. *Child Development, 59*, 897–907.
Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5, Serial No. 242).
Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language, 44*, 677–694.
Frausel, R. R., Richland, L. E., Levine, S. C., & Goldin-Meadow, S. (2021). Personal narrative as a "breeding ground" for higher-order thinking talk in early parent–child interactions. *Developmental Psychology, 57*, 519–534.
Frausel, R. R., Silvey, C., Freeman, C., Dowling, N., Richland, L. E., Levine, S. C., ... Goldin-Meadow, S. (2020). The origins of higher-order thinking lie in children's spontaneous talk across the pre-school years. *Cognition, 200* 104274.
Fulkerson, A. L., & Waxman, S. R. (2007). Words (but not tones) facilitate object categorization: Evidence from 6- and 12-month-olds. *Cognition, 105*, 218–228.
Gainotti, G. (2014). Old and recent approaches to the problem of non-verbal conceptual disorders in aphasic patients. *Cortex, 53*, 78–89.
Gathercole, S. E., Willis, C. S., Emslie, H., & Baddeley, A. D. (1992). Phonological memory and vocabulary development during the early school years: A longitudinal study. *Developmental Psychology, 28*, 887–898.
Gelman, A., & Carlin, J. (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science, 9*, 641–651.
Gelman, S. A., & Davidson, N. S. (2013). Conceptual influences on category-based induction. *Cognitive Psychology, 66*, 327–353.
Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive Development, 14*, 487–513.
Gleitman, L., & Fisher, C. (2005). Universal aspects of word learning. In J. McGilvray (Ed.), *The Cambridge companion to Chomsky* (pp. 123–142). Cambridge, UK: Cambridge University Press.
Graham, S. A., Booth, A. E., & Waxman, S. R. (2012). Words are not merely features: Only consistently applied nouns guide 4-year-olds' inferences about object categories. *Language Learning and Development, 8*, 136–145.
Hart, S. A., Petrill, S. A., Deckard, K. D., & Thompson, L. A. (2007). SES and CHAOS as environmental mediators of cognitive ability: A longitudinal genetic analysis. *Intelligence, 35*, 233–242.
Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science, 20*, 729–739.
Huttenlocher, J., Waterfall, H., Vasilyeva, M., Vevea, J., & Hedges, L. V. (2010). Sources of variability in children's language growth. *Cognitive Psychology, 61*, 343–365.
Jensen, A. R. (1998). *The g factor: The science of mental ability*. Westport, CT: Praeger.
Jones, A., Atkinson, J., Marshall, C., Botting, N., St Clair, M. C., & Morgan, G. (2020). Expressive vocabulary predicts nonverbal executive function: A 2-year longitudinal study of deaf and hearing children. *Child Development, 91*, e400–e414.
Kelly, D. J. (1998). A clinical synthesis of the "late talker" literature: Implications for service delivery. *Language, Speech, and Hearing Services in Schools, 29*(2), 76–84.
Kievit, R. A. (2020). Sensitive periods in cognitive development: A mutualistic perspective. *Current Opinion in Behavioral Sciences, 36*, 144–149.
Kievit, R. A., Hofman, A., & Nation, K. (2019). Mutualistic coupling between vocabulary and reasoning in young children: A replication and extension of the study by Kievit et al. (2017). *Psychological Science, 30*, 1245–1252.
Kievit, R. A., Lindenberger, U., Goodyer, I. M., Jones, P. B., Fonagy, P., Bullmore, E. T., & Dolan, R. J. (2017). Mutualistic coupling between vocabulary and reasoning supports cognitive development during late adolescence and early adulthood. *Psychological Science, 28*, 1419–1431.
Kuhn, L. J., Willoughby, M. T., Vernon-Feagans, L., & Blair, C. B. (2016). The contribution of children's time-specific and longitudinal expressive language skills on developmental trajectories of executive function. *Journal of Experimental Child Psychology, 148*, 20–34.
Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods, 44*, 978–990.
LeFevre, J.-A., Fast, L., Skwarchuk, S.-L., Smith-Chant, B. L., Bisanz, J., Kamawar, D., & Penner-Wilger, M. (2010). Pathways to mathematics: Longitudinal predictors of performance. *Child Development, 81*, 1753–1767.
Lewis, M., Colunga, E., & Lupyan, G. (2021). Superordinate word knowledge predicts longitudinal vocabulary growth. *Proceedings of the 43rd Annual Meeting of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
Lupyan, G. (2009). Extracommunicative functions of language: Verbal interference causes selective categorization impairments. *Psychonomic Bulletin & Review, 16*, 711–718.
Lupyan, G. (2012). What do words do? Towards a theory of language-augmented thought. In B. H. Ross (Ed.). *The psychology of learning and motivation* (Vol. 57, pp. 255–297). Cambridge, MA: Academic Press.
Lupyan, G. (2016). The centrality of language in human cognition. *Language Learning, 66*, 516–553.
Lupyan, G., & Mirman, D. (2013). Linking language and categorization: Evidence from aphasia. *Cortex, 49*, 1187–1194.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: Labels facilitate learning of novel categories. *Psychological Science, 18*, 1077–1082.
Lupyan, G., & Thompson-Schill, S. L. (2012). The evocative power of words: Activation of concepts by verbal and nonverbal means. *Journal of Experimental Psychology: General, 141*, 170–186.
MacWhinney, B. (2000). *The CHILDES project: The database* (Vol. 2) Hove, UK: Psychology Press.
Mervis, C. B., & Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development, 53*, 258–266.
Miller, G. A. (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
Miller, H. E., Vlach, H. A., & Simmering, V. R. (2017). Producing spatial words is not enough: Understanding the relation between language and spatial cognition. *Child Development, 88*, 1966–1982.
Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development, 78*, 622–646.
Morgan, P. L., Farkas, G., Hillemeier, M. M., Hammer, C. S., & Maczuga, S. (2015). 24-Month-old children with larger oral vocabularies display greater academic and behavioral functioning at kindergarten entry. *Child Development, 86*, 1351–1370.
Nazzi, T., & Gopnik, A. (2001). Linguistic and cognitive abilities in infancy: When does language become a tool for categorization? *Cognition, 80*, B11–B20.
Neuman, S. B., Newman, E. H., & Dwyer, J. (2011). Educational effects of a vocabulary intervention on preschoolers' word knowledge and conceptual development: A cluster-randomized trial. *Reading Research Quarterly, 46*, 249–272.
Newman, R. S., Rowe, M. L., & Ratner, N. B. (2016). Input and uptake at 7 months predicts toddler vocabulary: The role of child-directed speech and infant processing skills in language development. *Journal of Child Language, 43*, 1158–1173.
Peng, P., Lin, X., Ünal, Z. E., Lee, K., Namkung, J., Chow, J., & Sales, A. (2020). Examining the mutual relations between language and mathematics: A meta-analysis. *Psychological Bulletin, 146*, 595–634.
Perry, L. K., & Samuelson, L. K. (2013). The role of verbal labels in attention to dimensional similarity. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuch (Eds.), *Proceedings of the Thirty-Fifth Annual Conference of the Cognitive Science Society* (pp. 3217–3222). Austin, TX: Cognitive Science Society.
Plunkett, K., Hu, J.-F., & Cohen, L. B. (2008). Labels can override perceptual categories in early infancy. *Cognition, 106*, 665–681.
Ritchie, S. J., Bates, T. C., & Plomin, R. (2015). Does learning to read improve intelligence? A longitudinal multivariate analysis in identical twins from age 7 to 16. *Child Development, 86*, 23–36.
Roid, G. H., & Pomplun, M. (2012). *The Stanford-Binet Intelligence Scales* (5th ed.). New York: Guilford.
Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology, 8*, 382–439.
Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods, 51*, 1928–1941.
Simms, N. K., & Gentner, D. (2019). Finding the middle: Spatial language and spatial reasoning. *Cognitive Development, 50*, 177–194.
Simms, N. K., & Richland, L. E. (2019). Generating relations elicits a relational mindset in children. *Cognitive Science, 43* e12795.
Sloutsky, V. M., Lo, Y.-F., & Fisher, A. V. (2001). How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference. *Child Development, 72*, 1695–1709.
Snedeker, J., & Gleitman, L. (2004). Why is it hard to label our concepts? In D. G. Hall & S. R. Waxman (Eds.), *Weaving a lexicon* (illustrated ed., pp. 257–294). Cambridge, MA: MIT Press.
Socher, M., Ingebrand, E., Wass, M., & Lyxell, B. (2020). The relationship between reasoning and language ability: Comparing children with cochlear implants and children with typical hearing. *Logopedics, Phoniatrics, Vocology*. Advance online publication. doi:10.1080/14015439.2020.1834613
Song, J. Y., Demuth, K., & Morgan, J. (2018). Input and processing factors affecting infants' vocabulary size at 19 and 25 months. *Frontiers in Psychology, 9*. https://doi.org/10.3389/fpsyg.2018.02398.
Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science, 29*, 41–78.
Timm, A., Gelman, A., & Carlin, J. (2019). *retrodesign: Tools for Type S (sign) and Type M (magnitude) errors* (Version 0.1.0) [computer software]. https://CRAN.R-project.org/package=retrodesign
van der Maas, H. L. J., Dolan, C. V., Grasman, R. P. P. P., Wicherts, J. M., Huizenga, H. M., & Raijmakers, M. E. J. (2006). A dynamical model of general intelligence: The positive manifold of intelligence by mutualism. *Psychological Review, 113*, 842–861.
Vanluydt, E., Supply, A.-S., Verschaffel, L., & Van Dooren, W. (2021). The importance of specific mathematical language for early proportional reasoning. *Early Childhood Research Quarterly, 55*, 193–200.
Waxman, S. R. (2003). Links between object categorization and naming: Origins and emergence in human infants. In D. H. Rakison & L. M. Oakes (Eds.), *Early category and concept development: Making sense of the blooming, buzzing confusion* (pp. 213–241). Oxford, UK: Oxford University Press.
Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock-Johnson Tests of Cognitive Ability*. Allen, TX: DLM Teaching Resources.
Yu, C.-P., Maxfield, J. T., & Zelinsky, G. J. (2016). Searching for category-consistent features: A computational approach to understanding visual category representation. *Psychological Science, 27*, 870–884.
Zettersten, M., & Lupyan, G. (2020). Finding categories through words: More nameable features improve category learning. *Cognition, 196* 104135.