Decoupling literalist behavior from children's early metaphor comprehension abilities

Mary Beth Neff and Ingrid Lossius Falkum
University of Oslo, Norway

Abstract: Children's literalist responses to metaphor comprehension tasks are often taken to indicate deficient metaphor comprehension. We aimed to decouple this assumed (literalist) performance–(metaphor) competence link and investigate whether children's observed literalism is best explained by an early difficulty with metaphor. We assessed 3- to 7-year-olds' metaphor comprehension abilities using different novel functional, attributional, and psychological metaphors in a between-subjects design. We found that when not provided with literal options, children could derive metaphorical interpretations successfully. This was further supported by longer reaction times for metaphorical over literal interpretations. However, when literal options were available, even adults predominantly chose them over metaphorical interpretations. These findings challenge the view that children's literalism stems solely from difficulty with metaphor and urge researchers to more clearly distinguish studies assessing sensitivity to metaphorical meaning from those investigating the ability to prioritize a metaphorical interpretation over a literal one.

Keywords: metaphor; children; literalism; experimental; pragmatics

Corresponding author(s): Mary Beth Neff, Department of Philosophy, Classics, History of Art and Ideas, Blindernveien 31 Georg Morgenstiernes hus 0313 Oslo, Norway. Email: marybeth-neff@gmail.com

ORCID ID(s): Mary Beth Neff: https://orcid.org/0000-0002-9549-5936; Ingrid Lossius Falkum: https://orcid.org/0000-0002-1203-8036

Citation: Neff, M.B., & Falkum, I. L. (2025). Decoupling literalist behavior from children's early metaphor comprehension abilities. *Language Development Research, 5*(2), 130–171. http://doi.org/10.34842/ldr2025-819

Introduction

If you were to stub your toe and say, "Ah - my toe is on fire!", there are at least two possible interpretations to take—that your toe is literally on fire or that your toe is in a lot of pain. Given the context, adults would likely infer that you were in a lot of pain. Children, however, may take you to mean that your toe was actually on fire. We explored why, between these alternatives, children often favor literal interpretations. Specifically, we investigated whether children's preference for literal interpretations is best explained by early difficulties with understanding novel metaphors, as often suggested in previous research.

Recent evidence has suggested that children as young as 3 years old can make sense of novel metaphorical statements (5et al., 2024; Pouscoulous & Tomasello, 2020). However, despite this early ability, children often favor literal interpretations, even when the context does not support them (e.g., Martín-González et al., 2025) and sometimes up until preadolescence (e.g., Van Herwegen et al., 2013). This observed literalist preference among young children has often been referred to as their "literal bias," a concept that has been a long-standing issue in research on metaphor comprehension development (see Vosniadou, 1987 and Winner, 1988/1997 for reviews). More generally, children's fragile performance on pragmatic tasks involving nonliteral uses of language (e.g., metaphor, irony, implicature) is puzzling in light of the growing body of evidence showing children's early pragmatic sophistication in other domains, including prelinguistic communication, word learning, and referential communication (see Matthews, 2014 and Zufferey, 2015 for reviews).

In early studies, researchers presumed children's literalist tendencies reflected a developmental period in which children are incapable of accessing metaphoric meaning until they reach preadolescence (so called metaphor deficit or literal stage accounts; see Vosniadou, 1987 and Winner, 1988/1997 for reviews). Researchers with this view claimed children moved from an initial creative and flexible stage—observed most notably in studies of children's early metaphor productions (e.g., Billow, 1981; Winner et al., 1980)—toward a strictly literal stage that persists throughout their preschool years (Gardner et al., 1975; Levorato & Cacciari, 2002). This idea of a fixed stage has since been criticized, as children have shown some capacity for figurative language within this supposed literal period (e.g., Di Paola et al., 2020; Gentner, 1977; Gottfried, 1997; Pouscoulous & Tomasello, 2020; Özçalişkan, 2005). However, while literal stage accounts have fallen out of favor, it is still commonly assumed that literalist tendencies reflect a difficulty with figurative language comprehension (e.g., Cacciari & Padovani, 2012; Long et al., 2021; Vosniadou et al., 1984; Winner, 1988/1997; Winner et al., 1976).

In light of other cases where children's competencies had previously been underdemonstrated in experimental settings (e.g., scalar implicature; Horowitz et al., 2018; Skordos & Papafragou, 2016), more recent research has begun to explore whether children's past difficulty with metaphor could be explained by task difficulty instead of a deficit in metaphor comprehension (Colston, 2020; Di Paola et al., 2020; Pouscoulous, 2011; Vosniadou et al., 1984). For example, previous tasks often relied on a sophisticated verbal reasoning ability where children had to explain what the experimenters really meant in context to demonstrate metaphor comprehension (for reviews, see Pouscoulous, 2011 and Vosniadou, 1987). The ability to comprehend is distinct from the ability to succinctly verbalize what is comprehended, and even adults struggle with this discrepancy in experimental settings (Faitaki & Murphy, 2019). Now, instead of using complicated verbal reasoning tasks, recent metaphor comprehension tasks typically use forced-choice paradigms that require children to disambiguate between literal and metaphorical interpretations (e.g., Di Paola et al., 2020; Long et al., 2021; Van Herwegen et al., 2013). For example, in a study with 5-year-olds, researchers asked children to pair one of three images with the statement "Lucy is a parrot" from either a literal depiction of a parrot, a metaphorical depiction of a girl resembling a parrot, or a distractor girl (Long et al., 2021). The use of these forced-choice style tasks often corresponds with findings showing an earlier onset of metaphor comprehension (e.g., Almohammadi et al., 2025; Pouscoulous & Tomasello, 2020). However, a preference for literal interpretations still appears in many studies on young children's metaphor comprehension (Vicente & Falkum, 2021; Winner, 1988/1997). For example, in Long et al. (2021) younger children chose literally ~80% of the time (see also Martín-González et al., 2025 for similar results among older children).

Because these newer forced-choice tasks are assumed to better scaffold children's developing abilities, any error children make on them tends to be taken as even stronger evidence against robust metaphor comprehension. In fact, presenting children with both literal and metaphorical interpretations in these forced-choice scenarios stems from the assumption that proving children's ability to comprehend metaphorical meaning requires demonstrating their capacity to ignore literal alternatives (Winner, 1988/1997). In these cases, using literal competitors is meant to act as a more rigorous test of comprehension because it shows that participants can ignore a highly competing alternative. However, we argue that while pitting literal competitors against metaphorical ones is a common practice, it makes the task about why children prioritize one interpretation over another rather than a test of sensitivity to metaphorical meaning alone.

However, not only has this (literalist)performance–(metaphor)competence relationship never been tested but inferring children's metaphor comprehension from their literal responses assumes literalist responding is inherently incorrect. These interpretations are reasonable if we are to assume that the literal interpretation is just a distractor, or that the metaphorical interpretations are the only correct choices. However, because both literal and metaphorical interpretations are made explicitly available in these tasks (usually via visual evidence), the 'correct' answer is in fact, ambiguous. This ambiguity makes it difficult to determine whether children's literalism is due to an inability to reason metaphorically or whether children perhaps privilege literal interpretations when they are available.

Long and colleagues (2021) also commented on the ambiguity of these literal versus metaphor disambiguation tasks (see also Gardner & Winner, 1978). In their comparison study with 13-year-olds (an age at which previous research has asserted children should be sensitive to metaphorical meaning, Vosniadou, 1987; Willinger et al., 2017; Winner, 1988/1997), they found that older children performed at chance and were not strictly metaphor-biased. They concluded that older children's sensitivity to both literal and metaphorical meanings likely made them sensitive to the ambiguity of the task as well—compared to the 5-year-olds whose literally biased performance suggested sensitivity to the literal meaning only. It could be that younger children's more apparent literalist preferences reflect a lack of sensitivity to metaphorical meaning; however, the fact that there *was* ambiguity makes this difficult to determine.

Despite the literature moving away from literal stage (or metaphor deficit) wording, children's errors—most of which are literal responses—are still presumed to reflect a difficulty in understanding metaphors. Regardless of whether this link between literalist performance and difficulty with metaphor exists, if the aim of metaphor comprehension tasks is to see if children can derive metaphorical interpretations, then including a highly competing literal interpretation complicates that goal. Decoupling these literal and metaphorical choices may therefore offer a more sensitive measure of early metaphor comprehension.

In a recent study, Pouscoulous and Tomasello (2020) replaced literal alternatives with distractors and indeed found that 3-year-olds could comprehend novel, perceptually based metaphors. For example, children were given two toy cars—one with a large sack on its roof (metaphorical) and one with a similar sack inside (distractor)—and asked to "Pick the car with the backpack." Children chose metaphorical depictions over distractors, demonstrating that even 3-year-olds can access metaphorical meaning in a minimal linguistic context.

Given the participants' comparatively young age in this study, their success in the task may also support a possible distinction between children's literalist tendencies and their ability to understand metaphor. However, because Pouscoulous and Tomasello (2020) did not provide literal options, it is unclear whether children's success should be attributed to the type of metaphor tested or the absence of literal interpretation options. It has been suggested that visual or attributional metaphors, like those used in Pouscoulous and Tomasello, are more accessible and easier to comprehend (Gentner, 1977; Winner et al., 1976), so it could have been that children's success was limited to those specific items. Additionally, because children appear to understand these more straightforward attributional metaphors, they may have been less likely to select literal alternatives had such options been provided. Since the task did not include literal options, it remains unclear what underpins children's success. Consequently, it cannot yet be determined whether literal responses reflect a deficient metaphor comprehension.

In the current study, we sought to unpack this coupling between literalist responses and metaphor comprehension by replicating and extending Pouscoulous and Tomasello (2020) to include more abstract metaphor types and literal competitors. We tested 3- to 7-year-olds to see if they could understand novel metaphors in the absence of literal options (Experiment 1) and whether the presence of literal options affected performance (Experiment 2). In Experiment 1, children chose which image, between a metaphorical and a distractor, matches a metaphoric statement. In Experiment 2, we tested different children on the same four sets of metaphoric statements and replaced the distractor images with literal depictions. We also ran a third experiment in which we gave children the literal images from Experiment 2 and the distractor images from Experiment 1 to get a sense of the different age groups' understanding of the test items and to compare possible processing differences across the experiments.

If children could make sense of metaphoric statements in the first experiment but performed at chance or responded literally in the second, that would suggest that metaphorical meaning may be accessible to children, but the presence of literal competitors could mask their early reasoning abilities in ambiguous settings. This distinction between performance in Experiments 1 and 2 would also support previous work demonstrating that children can grasp metaphors early on and challenge the idea that literalism results from protracted metaphor comprehension development.

In addition to our original question, we explored how children progress through the tasks using their reaction time. We also looked for any developmental patterns or differences in children's understanding of the different types of metaphors.

Previous research suggests a linear development of metaphor type (Asch & Nerlove, 1960; Gentner, 1977; Winner et al., 1976)—where children understand functional metaphors before psychological metaphors but both later than attributional metaphors[1]—so it could be that children's overall task performance improves with age. However, additional research has found a U-shaped curve in some nonliteral language development, a possibility that might extend to metaphor comprehension as well (Gardner et al., 1975; see also Köder & Falkum, 2020 for an example in children's metonymy acquisition). We extended the items to include these other metaphor types to provide a more robust test of metaphor comprehension. However, any predictions of metaphor type, age, and their interactions were purely exploratory, as neither these metaphor types nor their interactions with age have been tested systematically under this decoupled lens.

Our reaction time predictions were equally speculative, as few reaction time measures existed in metaphor tasks using developmental populations. However, using a reaction-time-as-processing-effort approach (as in Di Paola et al., 2020), we assumed the following. If children in Experiment 2 responded more slowly than those in Experiment 3, it would indicate that they were still sensitive to the competing options in Experiment 2, even if their responses were equally literal. Conversely, if their response times in Experiment 2 were comparable to those in Experiment 3, it would suggest that the children did not perceive or engage with the ambiguity of the experimental context. If children chose metaphorical interpretations above chance in Experiment 1 and took longer to respond compared to children who chose literal interpretations in Experiment 3, this would align with findings suggesting that metaphorical interpretations are more costly than literal interpretations (as in Noveck et al., 2001).

Experiment 1

To investigate children's early metaphor comprehension abilities without the presence of a competing literal interpretation, we first replicated the study by Pouscoulous and Tomasello (2020) and extended it to include novel concrete (i.e., attributional) and abstract (i.e., functional and psychological) metaphors and children from 3 to 7 years old. As in the original study, we expected children to pick metaphorical depictions above chance by eliminating the conflicting literal information. However, if children could not understand the metaphors, we expected them to perform at chance because there should have been no way for them to reconcile the differences between metaphor and distractor images if they were not capable of overcoming literal meaning.

1 Terminology for these metaphors varies (e.g., functional and psychological often are referred to as "abstract" metaphors and attributional as "concrete" or "perceptual"). For the purpose of this study, functional metaphors involve metaphorical relations derived from an object's function, whereas psychological metaphors involve relations to internal states (see Table 1 for a list of items).

In Pouscoulous and Tomasello (2020), 3-year-olds performed nearly at ceiling on attributional metaphor understanding. We expected to replicate this performance on trials adapting their original stimuli. However, as research on metaphor comprehension rarely tests multiple types of metaphors and age groups, we also explored the interaction between metaphor type and age to isolate further whether comprehension difficulties could be specific to different types of metaphors at specific developmental periods. All experimental protocols adhered to The National Committee for Research Ethics in the Social Sciences and the Humanities guidelines and have been approved by the Norwegian Agency for Shared Services in Education and Research (Reference Number: 596365). The study was also preregistered (https://osf.io/vauw2), and all materials, data, and analysis scripts are available on our OSF project page (https://osf.io/jkq9w/). In the next section, we describe ways our study differs from the original and the rationale.

Method

Model Selection and Sampling Plan

Prior to running the study, we ran full and null model simulations to assess the probability of successful model convergence (n = 120; for 1000 simulations) and confirmed model feasibility. For the complex model, including all correlational parameters, the probability of successful model convergence was 0.965; for the simple model excluding those parameters, the probability was 1. For each age group, planned sample size estimates were 20–26 for experimental conditions and 10–15 for the Experiment 3. Minimum estimates were informed by our model feasibility simulation (with total n = 120 for experimental conditions) and maximums by the reported samples in Pouscoulous and Tomasello (2020), referencing both final ($N_{\text{experiment1-2}} = 24$; $N_{\text{experiment3}} = 12$) and excluded ($N = 5$) samples. Supplementary analysis information for this project appears on our OSF page (https://osf.io/jkq9w/).

We chose to run our studies between subjects because we were interested in whether children could generally access metaphorical meaning and if they preferred literal meaning when it was available. Additionally, because we propose that providing literal alternatives may be problematic for testing metaphor understanding, we are less interested in exploring children's individual literal biases in the context of these paradigms. However, to mitigate concerns regarding potential individual variances across samples, we tested experiments across each testing session. We also assessed children's inhibitory control abilities using the DayNight task (Gerstadt et al., 1994) and collected language and demographic information from guardians, adapting the protocol from The ManyBabies Consortium (2020) to create a profile for each group of participants. Profiles were largely similar across samples and appear in the Supplementary Information section of this paper.

Participants and Design

We tested 82 3- to 7-year-old Norwegian-speaking, typically developing children (28 3-year-olds, 27 5-year-olds, and 27 7-year-olds). Of those tested, three children withdrew (1 per age group), and three were excluded (all 3-year-olds: one for a technical error, one for sibling interference, and one for practice failure). Children were given the same 20 test items in a fully randomized order, with metaphor and picture location on the right or left counterbalanced across participants.

Materials

In the original study, Pouscoulous and Tomasello (2020) always constructed the metaphors from perceptual similarities relating to the human body. Previous research (e.g., Gentner, 1977) has shown that children tend to perform better with body-related metaphors, supposedly because visual comparisons and spatial relations to the human body (e.g., "head = top") are more intuitive and easier to grasp. To avoid having the items all share a specific relation, we broadened the metaphoric devices to include functional and psychological metaphors as well as created new attributional metaphors without this body relation.

Additionally, following concerns regarding contrastive inference confounds discussed in Pouscoulous and Tomasello (2020, pp. 164–166), we adapted two of the original items. For "The dog with the brown shoes," we changed the distractor image to a dog with brown ears instead of the original dog with a brown bow so that both referents contained multiple brown features. For "The bottle with the big belly," we omitted the word "big" to avoid children choosing pictures from size (i.e., adjective) cues alone.

We made our metaphors by first compiling metaphor stimuli used in previous research testing similar age groups. We then either refined ones taken from that list or devised additional novel metaphors, taking inspiration from literature on children's understanding of object functions and mental state language (e.g., Bloom, 2001; Callanan et al., 2007; Deák, 2006; Estes et al., 1989; Harris et al., 2005). We also confirmed that all vocabulary used to construct the test referents were generally acquired by typically developing, Norwegian-speaking 3-year-olds. We did this by referencing the Norwegian vocabulary database, Ordforrådet (Lind et al., 2013; Simonsen et al., 2014), and the Stanford Wordbank (Frank et al., 2016). Additionally, we ran a control experiment in which we tested children on literal versus distractor images to ensure that children of similar ages could understand the test utterances without metaphorical constructions. See Experiment 3 later in this paper for this methodology and results.

Once we had a list of workable statements, we adapted them to the format used by Pouscoulous and Tomasello (2020) such that metaphors described a property of the subject of the utterance and were embedded in a referential statement (e.g., *Vis meg tårnet med hatten*/Show me the tower with the hat). After this process, we chose the final 20 metaphors based on how easily they could be depicted. Ten of these 20 were concrete (five from Pouscoulous and Tomasello and five new attributional items), and 10 were abstract (five novel functional and five novel psychological items).

For the images, we replicated the visual depictions from the Pouscoulous and Tomasello (2020) study near exactly (they are the cartoon equivalents of the physical toys used in the original study). All other images followed similar protocols to these items, where each image had to contain the referent and match on overall visual salience. We modeled the literal images on the distractor images and constructed the distractor images by mirroring the visual characteristics of the metaphorical images in a nonmetaphorical way. For example, in the phrase "The tree with the arms," the metaphorical tree had two branches that extended from the middle of the tree to look like arms. In contrast, the distractor tree had two roots that extended equally from the bottom of the tree, so they did not in any way relate to the metaphoric referent "arms." Table 1 presents the final 20 metaphoric statements and their English translations. Visual depictions appear on OSF (Løvstakken & Neff, 2024) though, see Figure 1 for an example.

Lastly, we ran a preference assessment test to account for potential preferences for one image over the other across sets. Similar to Pouscoulous and Tomasello (2020), we presented 3- to 7-year-old children either with metaphor and distractor items (n = 35), metaphor and literal items (n = 36), or literal and distractor items (n = 14), and we asked them to *Vis meg en* (Show me one). Children showed no preferences for either metaphor or literal over distractor images (p > .05). However, individual binomial tests confirmed significant preferences for a subset of images in the metaphor versus literal comparison group (of which, most were preferences for the metaphorical items over the literal ones). We made some adjustments following this finding, and preferences were removed (n = 20, p > .05). Raw data for these assessments, alongside original images, appears on OSF (https://osf.io/jkq9w/).

Table 1. ***Test sentences and their English translations***

Metaphor Type	Statement in Norwegian	Statement in English
Replicated attributional	Tårnet med hatten	The tower with the hat [a]
	Bilen med ryggsekken	The car with the backpack [a]
	Gulroten med håret	The carrot with the hair [a]
	Hunden med den brune sko	The dog with the brown shoes [a, b]
	Flasken med magen	The bottle with the belly [a, b]
New attributional	Treet med armene	The tree with the arms
	Hodet med spagettien	The head with the spaghetti
	Himmelen med kjeksen	The sky with the cookie
	Kakaoen med putene	The hot cocoa with the pillows
	Treet med de stekte eggene	The tree with the fried eggs
Functional	Frosken med paraplyen	The frog with the umbrella
	Reven med lommelykten	The fox with the flashlight
	Apen med hammeren	The monkey with the hammer
	Ekornet med koppen	The squirrel with the cup
	Larven på flyet	The bug on the plane
Psychological	Planten som er trist	The plant that is sad
	Ballene som er glade	The balls that are happy
	Tegningen som er sint	The drawing that is angry
	Baggen som sover	The bag that is sleeping
	Gutten som brenner	The boy that is on fire

[a] Items are an original subset from Pouscoulous and Tomasello (2020).
[b] Items changed to circumvent contrastive inferences.

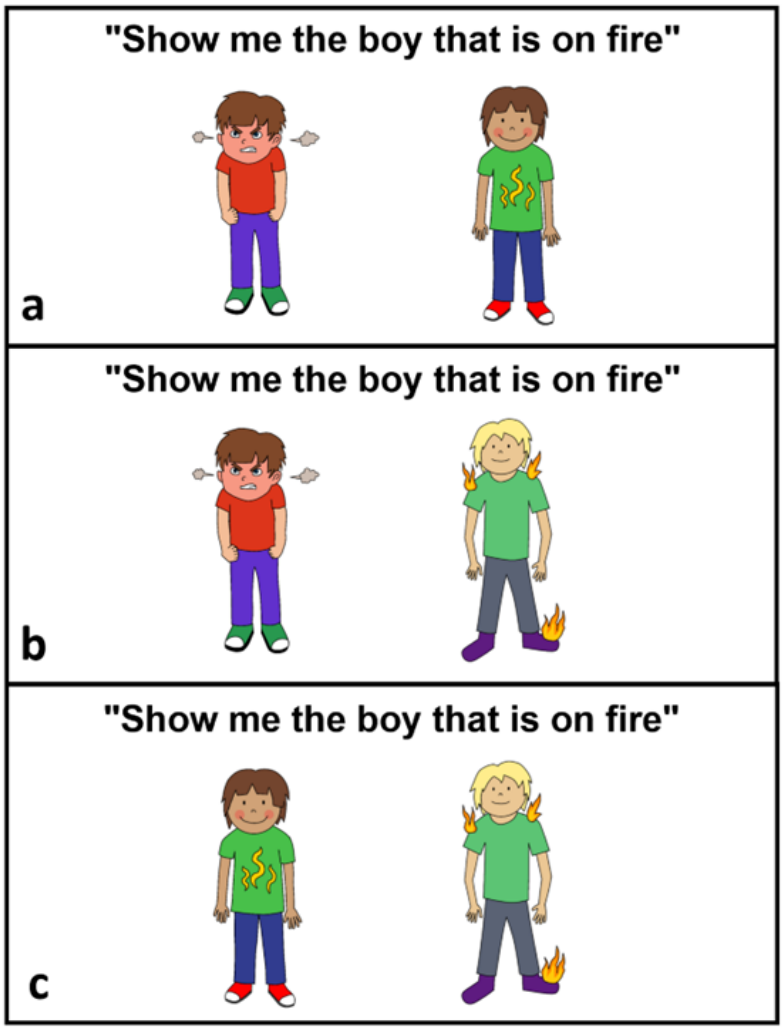

Figure 1. ***In child point of view: Schematic of test trials for Experiments 1 (a), 2 (b), and 3 (c).***
Note. All statements were aurally presented in Norwegian, and no text bubbles were used.

Procedure

Testing was done on tablets using the program PsychoPy/Pavlovia (Bridges et al., 2020; Peirce et al., 2019). The testing occurred in museums, kindergarten classrooms, and schools around Oslo. Testing always took place in a separate room to avoid other children and distractions. We only invited children whose guardians returned the signed consent form to participate. In addition to the guardians' informed consent, we made participating children aware of the voluntary nature of the study and all

children had to assent to participate. Children also received a small gift (e.g., an eraser or a diploma) for their participation.

To begin the task, an experimenter introduced the children to two characters on a tablet and explained that they would be playing a game where the characters would ask questions about different pictures on the screen. In the original Pouscoulous and Tomasello (2020) study, children interacted with two experimenters and real objects. We made these changes to familiarize the children with one character and play the game with another. We chose characters because it reduced the need to control for social information and allowed us to remove the other character from the environment when they were not speaking. This removal also reduced the amount of communicative style switching in the task (i.e., speaking literally and figuratively). Children interacted with both characters; however, character identity was fixed in the practice and test sessions. The experimenter was a trained research assistant and a native Norwegian speaker.

Practice Trials. To ensure children understood the game, the experimenter told them they would play a practice game with one of the characters first. After introducing the child to the two characters on the screen, one character left, and the other remained to take the child through four practice trials. To be included in the final experiment, children had to have answered three out of four practice trials correctly, as was the case in Pouscoulous and Tomasello (2020).
The practice trials were identical to the test trials; except they did not include metaphorical statements. For example, the practice character asked the child to "Pick the apple that is red" while presenting an image of a red apple and a green apple. We took two of the practice trials directly from Pouscoulous and Tomasello (2020), and we altered the remaining two to familiarize children with more pragmatic choices. In piloting, we noticed that using only literal practice trials trained children to expect a specific referential match (i.e., to look for a "correct" and literal response). In our experiments, the literalness varied, making the correctness less defined, so we created two 'pragmatic' practice trials to discourage this type of response heuristic. Pragmatic trials included referents that were not entirely satisfactory or prototypical, but one was always more appropriate than the other. For example, we asked children to pick "the line that is straight" and presented them with one nearly but not perfectly straight line and one squiggly line (Ronderos et al., 2022). Like in the original Pouscoulous and Tomasello trials, children got feedback on their responses.

Test Trials. After practice, the experimenter told the children they had finished practicing and would now begin to play the actual game with the other character from the introductions. In the test, the character from the familiarization left, and the other character from the introduction returned and told the child that they would also like to play. As in practice, two images (now a metaphor and a distractor image) for each statement appeared on the screen, and the character asked the child to select

one of the images that aligned with the instruction: "Show me the" plus a metaphoric statement. For example, the character said, "Vis meg tårnet med hatten" (Show me the tower with the hat) while an image of a tower with a red balcony (i.e., distractor image) and an image of another tower with a large red roof (i.e., metaphoric depiction) appeared. After the child selected an image, the subsequent trial immediately began. All test statements were metaphorical, and no feedback was provided.

The experiment session was audio recorded, and children's picture selections and reaction times were recorded directly via the tablet. Distractor image location, which we noted as left or right according to the child's point of view, was fixed in practice (i.e., ABBA) and counterbalanced in the test, such that half the metaphoric images appeared on the children's right and half on the left. Statement order was also fixed in practice but fully randomized in the test. See Figure 1 for a schematic of the procedure.

Results

Model Selection

Children's picture selections were analyzed using a generalized linear mixed model in R (lme4 package, Bates et al., 2015; R Core Team, 2021) using a binomial distribution with metaphoric responses coded as 1 and the distractor coded as 0. The full model included age group (z transformed), metaphor type, and their interactions as fixed factors; items and subject identity as random effects (Clark, 1973); and random slopes of age and metaphor type. The null model was identical but with metaphor type and its interaction with age removed. We made inferences regarding effects on picture selections via full-null model comparisons using an analysis of variance ($p \leq 0.05$) via the lmerTest R package (Kuznetsova et al., 2017).

In our preregistration (https://osf.io/vauw2) we outlined a confirmatory model that only included the effect of experiment type on picture selection, as well as an additional exploratory model that looked at potential interaction effects of age and metaphor type. We made these choices because, for our comparison across experiments, the model isolating experiment was more theoretically coherent. However, in isolating the effects of this first study, we decided to progress with the exploratory model as the maximal structure that included effects of metaphor type, age, and their interactions more appropriately fit our experimental design. See Barr et al. (2013) and Schielzeth and Forstmeier (2009) for further discussion regarding maximal slope structures. There were two additional changes from preregistration, namely that location ID and trial ID were removed from random effects structures as testing location was balanced and randomization was set to full random, so both were collapsed within subject identity.

Model Results

The full model (AIC = 871.99) did not provide a significantly better fit to the data compared to the null model: AIC = 873.01, $\chi^2(2) = 2.98$, $p = .225$, suggesting that age alone had a significant effect on the proportion of metaphorical picture selections but not metaphor type or their interactions. From visual inspection of the data (see Figure 2), we surmised that all children selected metaphorical pictures at above-chance levels (chance = .50) but that 3-year-old children chose slightly fewer metaphorical pictures compared to 5- and 7-year-olds. However, for a more detailed overview of metaphor type and individual item performance, see the Supplementary Information section for a graph depicting all four metaphor types alongside tables of mean picture selections for each item by age group.

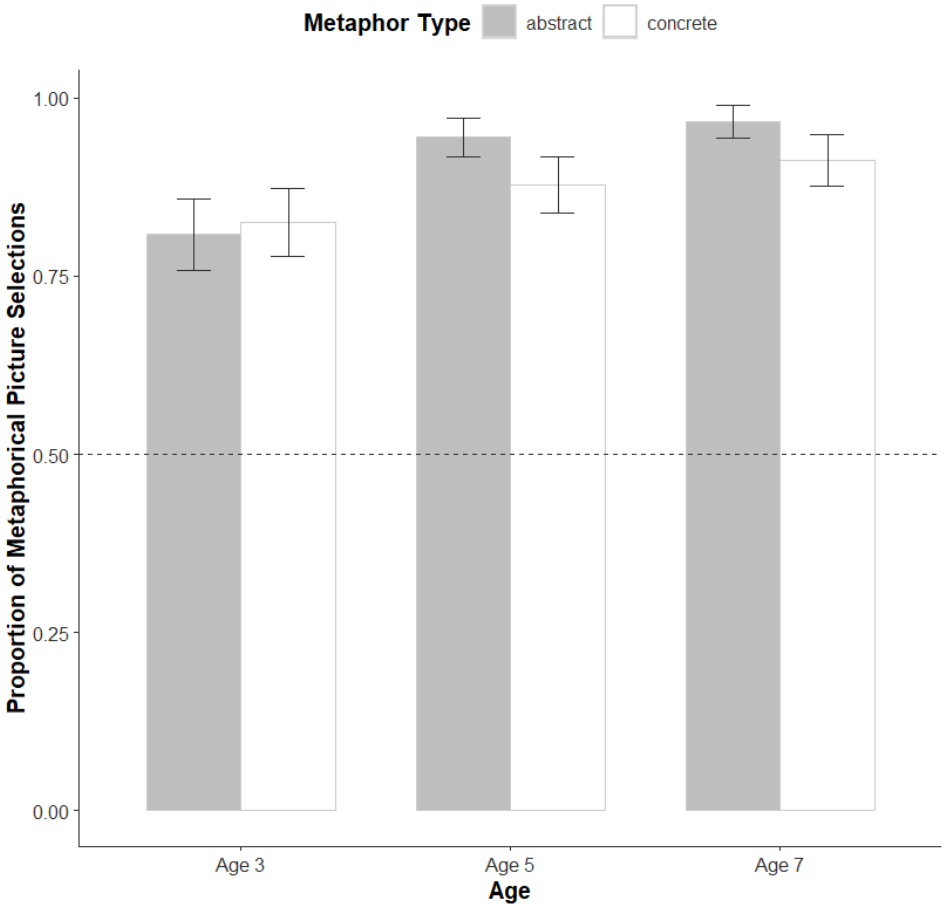

Figure 2. ***Proportion of metaphor selections in Experiment 1 by age and metaphor type.***

Experiment 2

Experiment 2 followed the exact procedure as Experiment 1—that is, we kept the metaphorical statement and depictions constant. However, in the test, we replaced the distractor images with literal depictions. We predicted reduced metaphoric selections in this second experiment because the literal interpretation was now available. For the predicted comparisons between experiments, if children's literalism reflected an early metaphor comprehension deficit, we would expect literal responding in Experiment 2, chance responding in Experiment 1, and for this overall pattern to be more apparent in our younger age groups. However, if we encountered literalist responding in Experiment 2 but above-chance metaphor selection in Experiment 1, this would contradict a direct mapping between literalism and protracted metaphor comprehension.

Method

Participants and Design

We tested 82 3- to 7-year-old Norwegian-speaking, typically developing children (29 3-year-olds, 27 5-year-olds, and 26 7-year-olds). Of those tested, one 3-year-old child withdrew, and three were excluded (two 3-year-olds and one 5-year-old for practice failures). Aside from replacing the distractor images with literal ones, we gave children the exact same 20 test items as Experiment 1 in a fully randomized order with metaphor picture location (i.e., right vs. left) counterbalanced across participants.

Results

Our model followed the same parameters as the first experiment, but to investigate the effect of experiment type on picture selection, we additionally included fixed effects of experiment type (including its interaction with metaphor type and age), as well as its random slope on metaphor type. The null model did not include the effect of experiment type.

The full model (AIC = 1,468.5) provided a significantly better fit to the data compared to the null model: AIC = 1,560.5, $\chi^2(4) = 100.04$, $p < 0.001$, meaning that including the effect of experiment type explained children's picture selections better than age alone. Visual inspection of the data (see Figure 3) showed that not only did children select fewer metaphorical images in Experiment 2, but they also selected them at below-chance levels, thus indicating a preference for the literal alternatives.

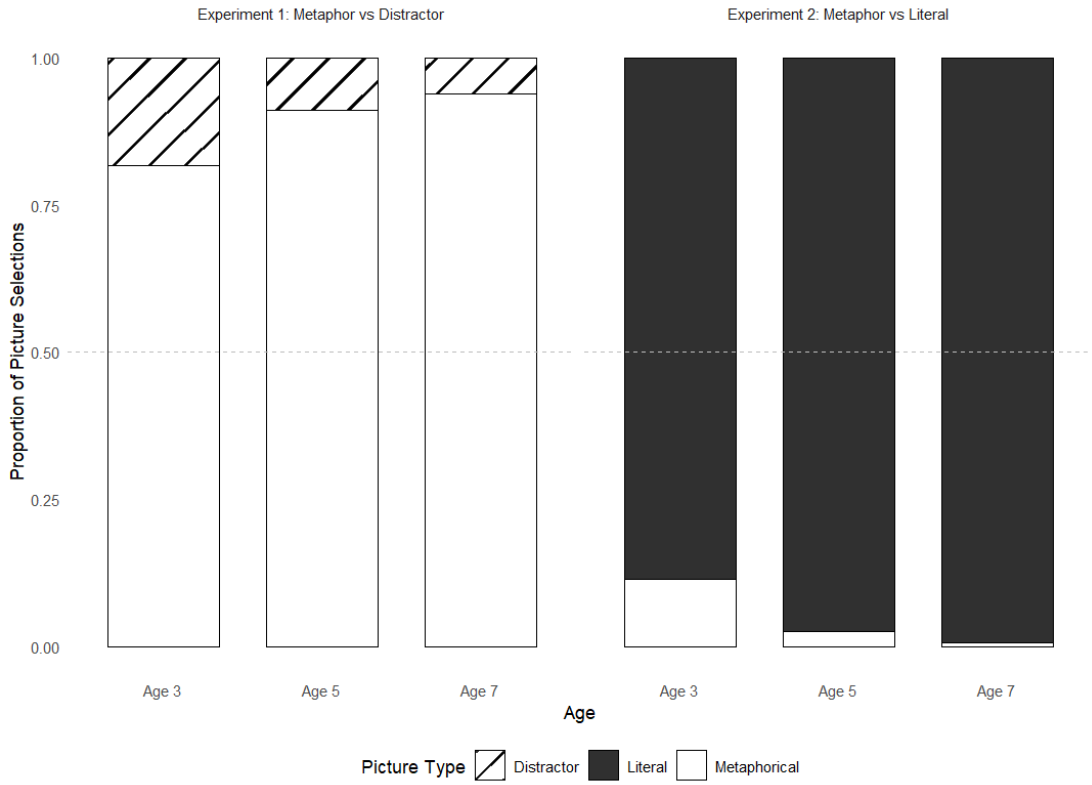

Figure 3. ***Proportion of picture selections in Experiments 1 and 2 with dotted line representing chance at 0.50.***

Experiment 3

To test whether children in these age ranges generally understood the test vocabulary, we ran a control experiment using the same procedure as Experiments 1 and 2. Although we constructed the items from words generally acquired by 30 months (Frank et al., 2016; Lind et al., 2013; Simonsen et al., 2014), we wanted a measure of children's relative understanding of our test vocabulary. We presented children in this study with the same utterances as in Experiments 1 and 2, but we did not give them metaphoric options. Instead, we presented the children with the distractor images used in Experiment 1 and the literal images used in Experiment 2. For example, for "the tree with the arms," the experimenter showed children a tree with cartoon arms (i.e., literal) and a different tree with its roots showing (i.e., distractor; see also Figure 1 for an example schematic). We expected children of all ages to choose literal items at above-chance levels, with older children responding at ceiling.

Additionally, in line with the exploratory reaction time data for Experiments 1 and 2, we planned to compare children's response times in Experiment 2 with these in Experiment 3 to determine whether children remained sensitive to the metaphorical interpretation in Experiment 2 despite their literalist responding. If this were the case, we expected to find faster responses in Experiment 3 compared to Experiment 2. Additionally, if children were slower in Experiment 1 compared to Experiment 3, it would suggest that children were pragmatically constructing metaphorical interpretations at an added cost in Experiment 1 compared to the literal interpretations they could access more directly in Experiment 3. Because our reaction time hypotheses were partly motivated by adult comparison research, we present these data in the Adult Experiment section to include all data in a single model and avoid multiple testing (see also Figure 5 for these data).

Method

Participants and Design

We tested 50 3- to 7-year-old Norwegian-speaking, typically developing children (17 3-year-olds, 16 5-year-olds, and 17 7-year-olds). Of those tested, three participants withdrew (one per age group), and we excluded two (one 3-year-old for practice failure and one 7-year-old for language requirements). Procedures followed the other experiments except that we set picture identities in the test to literal versus distractor images.

Results

Model selection followed previous experiments; however, we removed fixed effects of metaphor type and experiment. The confirmatory model, including fixed effects of age, explained the variance in the data better in the full model (AIC = 137.0) than in the null model. The null model just included a random intercept: AIC = 148.3, $\chi^2(3) = 17.296$, $p < 0.001$. Visual inspection of the data showed all participants selected literal pictures at above-chance levels; however, 5- and 7-year-olds selected target (i.e., literal) pictures slightly more often than 3-year-olds, but this difference was not significant ($p = 0.138$). See Figure 4 for all picture selection results across ages.

Adult Experiment

Children's metaphor comprehension is often assessed in terms of their ability to overcome a literal interpretation. We assert that pitting literal and metaphorical referents against one another adds additional challenges to these tasks and ultimately makes the goal of these tasks more ambiguous. Therefore, we sought to disentangle these response patterns to determine whether literalist response patterns are best explained by an early difficulty with metaphor.

We found in Experiment 1 that when there were no competing literal alternatives, children were able to select metaphorical interpretations at well above-chance performance. Additionally, despite the general capacity for metaphorical reasoning at these age ranges, we found that children chose literally when given the option in Experiment 2. The above-chance performance on metaphor comprehension in Experiment 1 and the literally biased responses in Experiment 2 speak to a distinction between the two phenomena (i.e., between literalism and early metaphor comprehension abilities). The fact that the differences in successful metaphorical selections were relative to the presence or absence of a literal competitor, rather than an effect of age or metaphor type, also supports our previously asserted suspicions regarding the potential impact literal items have on performance. However, given the general skepticism surrounding young children's capacity for figurative language, we set out to replicate these findings in an adult sample.

Additionally, something that has motivated this research from the beginning is that even from our own perspective, we could not discern why literal competitors would be inherently incorrect in these settings where they had been made explicitly available. Upon hearing the statement "My toe is on fire," one may not immediately expect the literal interpretation, but why would you not endorse that interpretation once provided with visual evidence that it is indeed afforded in that communicative context (see also Winner et al., 1980 for related discussion). Long et al. (2021) argued that the reduced literalist performance among 13-year-olds indicated a sensitivity to both metaphorical and literal meanings and therefore the ambiguity of the test interaction, whereas the literalist responding among 5-year-olds showed a lacking sensitivity. However, given the absence of context in these settings, it seems no less rational for even adults to choose literal interpretations when this is likely to be the least effortful way to resolve the ambiguity, and also one that enables reference assignment (Wilson & Sperber, 2006).

Given the implications of an early metaphor comprehension deficit derived from children's literalist responding in these settings, we wanted to see how adults perform in these tasks where literal and metaphorical meanings are pitted against one another. Assuming an adult's sensitivity to metaphorical meaning is less fragile than that of a 3-year-old, if even adults choose literal meanings when given the same choice scenarios, then this would reinforce our argument for making a distinction between literalist performance and developing competence with metaphor.

Method

Participants and Design

We tested 67 Norwegian-speaking adults (26 in the following Experiment 1, 26 following Experiment 2, and 15 in Experiment 3). The adult tasks were nearly identical to the children's tasks with minor adaptations (e.g., shortening the intertrial intervals and changing the inhibitory control task to the Flanker task; Eriksen, 1995). We made these modifications so the adults could participate autonomously. All of the experimental materials appear on the OSF website (https://osf.io/8s2fm/). We did not collect demographic profiling information for adults, but their eligibility to participate was contingent on their self-reporting of Norwegian as a main language.

As we did with the children, we ran experiments in parallel, randomly assigning adults to one of the three experimental conditions (i.e., Experiments 1, 2, and 3) with the target picture's side counterbalanced and test item presentation fully randomized across participants. We recruited participants from museums, libraries, and universities in greater Oslo. All participants signed a written consent before participating, were made aware of the voluntary nature of the task and had the opportunity to opt-in to a prize draw for participating.

Results

As we were comparing adult data across experiments, we ran a full-null model comparison in which the full model included age, experiment, and their interaction terms and the null model only included age. Both random effects structures accounted for random slopes of age and experiment on each item and random intercept of subject ID and metaphor type. However, because we included all experiments in the model, we removed metaphor type as a fixed variable as there was no manipulation of metaphor in Experiment 3.

We found that the full model specifying the effect of experiment type (AIC = 1,496.1) better explained the rate of target picture selections (with metaphor coded as the target in Experiments 1 and 2 and literal coded as the target in Experiment 3) than the null model specifying age alone (AIC = 1,625.5, $\chi^2[4] = 137.42$, $p < 0.001$). This finding suggests that 3-year-olds tended to select fewer target pictures compared to older participants and that Experiment 2 negatively affected target picture selection rates for all subjects. Like children, adults similarly chose metaphorical pictures fewer times in Experiment 2 compared to those in Experiment 1. Adults also chose these metaphorical pictures at above-chance levels in Experiment 1 but below-chance levels in Experiment 2, with near-ceiling and floor rates, respectively ($M = 0.97$, $SD = 0.17$; $M = 0.02$, $SD = 0.16$).

However, there were potential multicollinearity issues with the full model as age had a high variance inflation factor value (VIF = 3.97). Though see Fox and Monette (1992), Harrison et al. (2018), and Quinn and Keough (2002) for discussions regarding issues with multicollinearity calculations and thresholds (especially with models specifying interaction effects). We did not consider the correlation an issue for our data (following Allison, 2012), but to see if it affected possible inferences, we reran the models where we dropped the fixed effect of age in the full model because the experiment was always our registered effect of interest. We included a random intercept model for the null and found comparable results (with full model AIC = 1,635.6 and null AIC = 1719.4, $p < 0.001$).

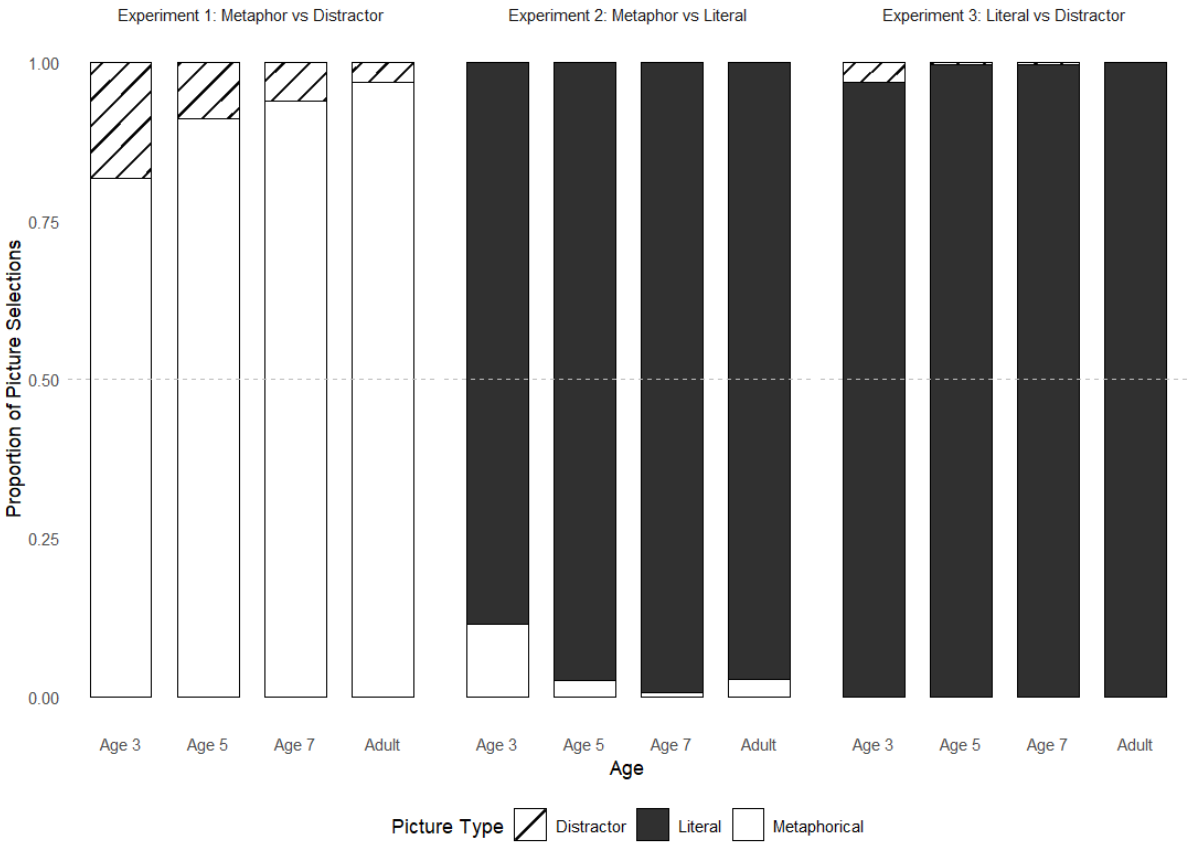

Figure 4. ***Proportion of picture selections by experiment and age with dotted line representing chance at 0.50.***

Exploratory Analysis of Reaction Time

Our primary aim in this paper was to disentangle literal and metaphorical response patterns. In doing so, we have also argued that directly contrasting literal and metaphorical items can obscure children's sensitivity to metaphorical meaning by introducing communicative ambiguity. Prior research has suggested that a shift from literal to metaphorical responses may reflect a developing sensitivity to ambiguity (Long et al., 2021). However, because most studies test metaphor comprehension by requiring children to override literal alternatives, it was difficult to determine at what point literal responses reflect a lack of sensitivity to metaphorical meaning versus interpretation difficulties due to communicative ambiguity. To look into these different factors, we conducted exploratory tests of reaction time to see whether there were any processing differences across literal choices.

We proposed that slower responses in Experiment 2 compared to Experiment 3 could reflect sensitivity to competing interpretations, even when responses were equally literal. In contrast, similar response times would suggest limited engagement with the ambiguity in Experiment 2. Additionally, if children in Experiment 1 selected metaphorical interpretations above chance and took longer than those giving literal responses in Experiment 3, this would support the view that deriving novel metaphorical meanings is more demanding (e.g., Noveck et al., 2001). However, while these analyses were planned from the outset, the specific predictions were speculative as our method departs from conventional approaches in the literature. Accordingly, we interpret any findings here with caution.

Results

We recorded reaction times from the offset of the test utterance to the picture selection and analyzed them using a linear mixed model mirroring the parameters explained in the picture selection analysis in the adult experiment. After fitting the model, we checked for normality by examining the Q-Q plot of residuals (Field, 2005), which we confirmed violated assumptions in a follow-up Shapiro test ($p < 0.001$). We then performed the Box-Cox procedure (Box & Cox, 1964) to find the appropriate transformation for our data ($\lambda = 0.02$) and reran all models using the transformed data which met model assumptions (Lo & Andrews, 2015).

The full model (AIC = 8,495.3) provided a better fit for the data compared to the null model (AIC = 8,756.1), indicating that participants' reaction times varied significantly based on the type of experiment and its interaction with age ($\chi^2[24] = 277.13$, $p < 0.001$). Generally, reaction times decreased with age. Specifically, 3-year-olds had the longest reaction times across experiments with Experiment 1 (M = 4.42 [3.97–4.87]), Experiment 2 (M = 2.64 [2.35–2.94]), and Experiment 3 (M = 1.85 [1.64–2.07]). However, the interaction between age and experiment had mixed effects. Upon visual inspection (see Figure 5), children took longer to respond in Experiment 1 compared to Experiments 2 and 3. In contrast, adults showed similar reaction times in Experiments 1 and 2, both of which were slower compared to Experiment 3. Only 7-year-olds responded with similar reaction times in Experiments 2 and 3, although 5-year-olds appeared to show a similar trend. However, 3- and 5-year-olds and adults seemingly took longer to make selections in Experiment 2 when compared to Experiment 3. We revisit these exploratory results in the General Discussion.

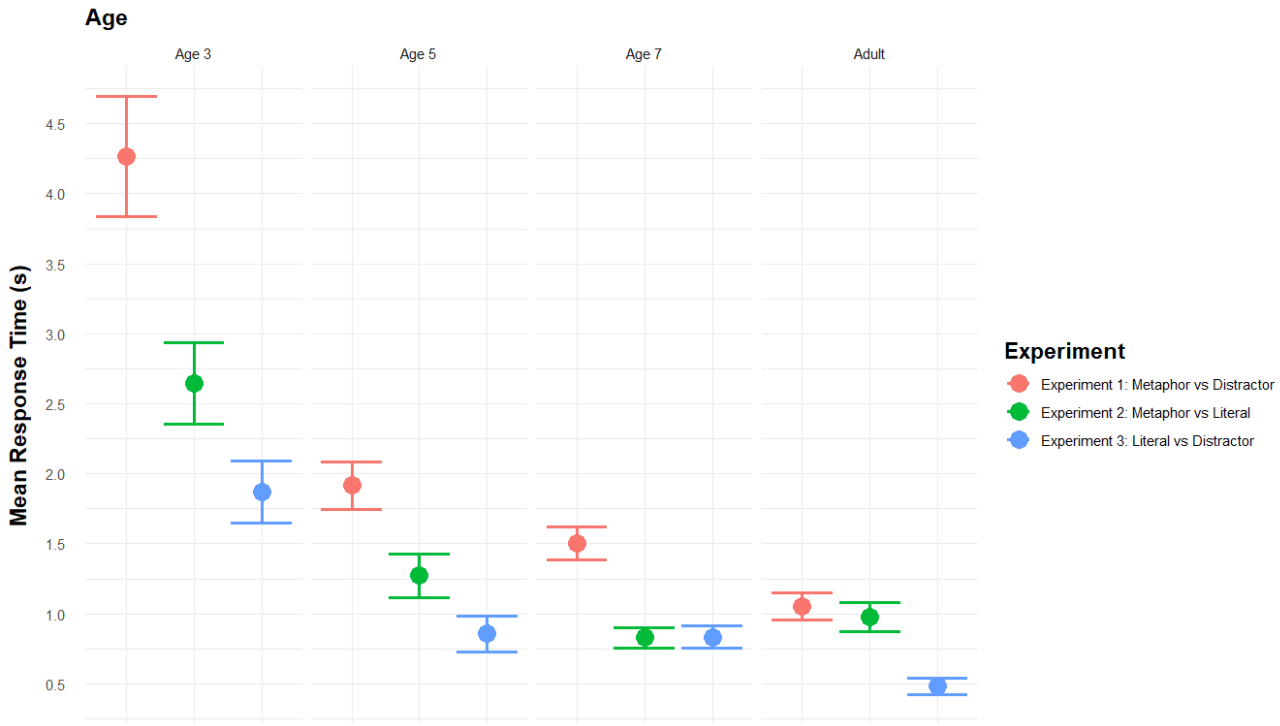

Figure 5. ***Mean response times in seconds observed across different age groups and experiment types with 95% confidence intervals.***

General Discussion

Children's literal interpretations were often assumed to depend on whether they could understand metaphors; however, this assumed link between literalism and early metaphor comprehension had never been tested directly. We investigated whether children's observed literalism in metaphor comprehension tasks is best explained by a developing sensitivity to metaphorical meaning using a between-subjects design. We found no support for claims that young children simply cannot understand metaphors, which is also in line with more recent findings (Almohammadi et al., 2025; Pouscoulous & Tomasello, 2020). Children from 3 years old were adept at our metaphor comprehension task when we removed the picture representing the literal interpretation from their response options. When we included literal options, younger children were not more literally biased than older children—which would follow from a metaphor deficit view, where children would become less literalist and more sensitive to figurative meanings with age. And even adults chose literal options exclusively in this context, where the target utterance was ambiguous between a metaphorical and a literal interpretation. Contrary to previous assumptions, young children's tendency toward literal interpretations does not appear to be explained by an underdeveloped sensitivity to figurative meaning.

Research on children's metaphor comprehension has often interpreted literal responses as evidence of limited sensitivity to metaphorical meaning. If this were the case, 3-year-olds—given their generally weaker performance on metaphor tasks (including in our Experiment 1)—should have shown the strongest literal bias in Experiment 2. Instead, they were no more likely than older children or adults to choose literal interpretations and, if anything, appeared slightly less biased. Adults' own tendency to prefer literal interpretations when both options were available further underscores that literalism cannot be straightforwardly equated with immature metaphor understanding. These findings challenge the assumption that literalist behavior declines linearly with age and call for a reevaluation of past research that equates literal responses with deficient metaphor competence.

One reinterpretation may be that literalist responses reflect children's preference for literal meanings, regardless of their understanding of metaphors, which consequently affects their performance on metaphor comprehension tasks. In tentative support of this claim, the presence of literal competitors, much like age, often predicts children's poor metaphor comprehension task performance (e.g., Van Herwegen et al., 2013; Vosniadou, 1989; Winner, 1988/1997). When given the opportunity in previous research, children chose literal interpretations (Vosniadou, 1987; Winner, 1988/1997), similar to what we found in our second experiment.

It was previously assumed that children chose literal options because they were unable to consider metaphorical alternatives. However, consistent with our findings, earlier studies that did not include literal options showed that children's responses were not random (Pouscoulous & Tomasello, 2020; Vosniadou, 1989; see Winner 1988/1997 for a review), nor did an ability with metaphor reduce literalist selections when a literal option was present (e.g., Long et al., 2021; Vosniadou, 1989). Instead, studies that did not include literal competitors often showed early sensitivity to metaphorical meaning (Pouscoulous & Tomasello, 2020; Vosniadou, 1989), and those that included literal competitors tended to find less metaphor-biased responding, even among older children (e.g., Long et al., 2021; Van Herwegen et al., 2013). Therefore, a misunderstanding of the tasks does not necessarily mark children's protracted metaphor comprehension development: children do not behave wholly randomly or appear confused; rather, they seem to demonstrate preferences for literal interpretations.

Relatedly, there are other factors that have been previously suggested to improve children's performance in metaphor comprehension tasks, which would also arguably help them overcome a literal preference. For example, Deamer (2013) found that children with better inhibitory control performed better on metaphor comprehension tasks. Traditionally, these findings have been argued to suggest that inhibitory control is necessary for children's metaphor comprehension because it allows them to handle conflict between literal and metaphorical meaning (Deamer, 2013; see also Pouscoulous, 2011 for a similar discussion). However, making this argument implies that literal meaning is inherently active, which children must overcome to appreciate metaphorical meaning—a processing account that has been debated (Gibbs & Colston, 2012). Additionally, in light of previous presumptions that young children could *not* access metaphorical meaning, it remains unclear how inhibitory control would aid children's sensitivity to metaphorical meaning because its relative absence would suggest the children would encounter no such conflict. Therefore, in keeping with a decoupled account, our data instead adds to these findings on inhibitory control by supporting the idea that literal meaning is salient (Deamer, 2013). As such, including these literal competitors introduces a nonarbitrary conflict in these tasks of metaphor comprehension and adds additional demands beyond testing sensitivity alone.

Although our findings suggest that literalist selections appear independent of metaphor comprehension, we cannot fully speak to the reasons behind children's early literalist tendencies. The goal of this paper was to disentangle metaphorical and literal responses from one another and not necessarily specify what underpins each individually in light of this decoupling. However, because children could successfully derive metaphorical interpretations in Experiment 1, and even adults chose literal options more often when available in Experiment 2, we assert that protracted metaphor comprehension is an unsatisfactory explanation.

While using literal alternatives remains a widely accepted method for testing metaphor comprehension, using literalism as evidence of a deficient development risks mischaracterizing both literal tendencies and underestimating children's metaphor comprehension abilities. Consequently, we caution against perpetuating the assumed mapping of literalist responding and (reduced) sensitivity to metaphorical meaning and suggest a need for more research into the underpinnings of literalist preferences independent from tests of figurative language comprehension. This is particularly relevant for research using literal competitors, as differences in our metaphorical picture selection and reaction time data across experiments showed that literal competitors add processing demands that researchers should consider when evaluating and making assumptions about early sensitivity to nonliteral meaning.

Methodological Considerations and Limitations

In our study, we replicated and adapted the choice paradigm used by Pouscoulous and Tomasello (2020). Though there have been discussions of potential confounds with this paradigm, such as contrastive inference abilities (e.g., Davies et al., 2021) and visual-associative processes (Pouscoulous & Tomasello 2020, pp. 164–166), the paradigm remains straightforward, has been successfully used in different populations (e.g., Almohammadi et al., 2025; Buehler et al., 2018), and, importantly, did not include literal competitors. However, to address these concerns, we adapted items that we felt could have been resolved via contrastive inferencing and still replicated and extended the original findings. We also broadened the metaphor types used to include more abstract metaphors and collected reaction time data to test the robustness of the task—particularly concerning its test of metaphor comprehension instead of visual associations.

In the original study, children could possibly have resolved the referential assignment via visual associations. For example, the word "hat" in the phrase "The tower with the hat" could have triggered the expectation that there would be something on top. Therefore, like Pouscoulous and Tomasello (2020), children may have selected the metaphorical option in our first experiment because it was the only option with a salient property on top. Although this is possible, it remains unclear how this resolution process would qualify as inherently nonfigurative or purely associative. One could argue that this process mirrors the one described for lexical broadening (Wilson & Carston, 2007), where the concept encoded by "the hat" is broadened to also denote other objects that occupy the same spatial relation of being on top to their bearers. Nonetheless, we also included more abstract metaphors to avoid these potential visual heuristics and found that children performed above chance on all metaphor types. Although this heuristic might be available in more attributional trials, it is unlikely to uniformly apply to our other metaphor types, where this meaning is more abstracted (e.g., as in "the drawing that is angry").

Also, because associative accounts are usually highlighted as a leaner contrast to a richer cognitive account, it remains to be explained why participants took generally longer to progress through the first experiment. Although contested as an account of how communication works, one could argue that the initial search for the literal meaning adds extra time (Grice, 1975; though cf. Sperber & Wilson, 2008). However, the fact that children took longer to respond literally in Experiment 2 compared to Experiment 3 contradicts this account because it suggests that children are not simply searching for literal meaning in the first instance. We also controlled for potential low-level features of the images that could contribute to how children responded. Our preference assessment data showed that children chose randomly between image comparisons (i.e., metaphorical vs. distractor, metaphorical vs. literal, and literal vs. distractor) when instructed with the neutral phrase "Show me one." Only when we made the metaphorical or literal meaning available in the main experiments did they reliably attend to one image over the other. Thus, regardless of broader concerns about what forced-choice tasks reveal about metaphor comprehension, our reaction time data and results across different metaphor types and test prompts suggest that children reliably selected specific referents based on the communicative context, rather than relying solely on visual shortcuts (though see Shanks, 2010, for a discussion on how associative and cognitive accounts can complement each other).

Lastly, our task did not use any disambiguating discourse context in line with previous research (though cf. Keil, 1986 and Gardner et al., 1975 for arguments against testing metaphor sans discourse). Because we did not have any disambiguating context, we cannot speak to the literal "bias" explicitly, as even though participants exclusively chose literally, they did not do so despite more appropriate alternatives (see Tversky & Kahneman, 1974 for discussions operationalizing biased responding). We do not see this as an issue with the current design, as one of our core motivations was to highlight potential issues with traditional forced-choice paradigms. Therefore, we intentionally chose these ambiguous scenarios to cohere with previous research. Additionally, although our findings show that literal competitors add nonarbitrary impact to tasks of metaphor comprehension, we do not claim to have resolved the literal bias full stop. However, considering that children could derive metaphorical interpretations without this prior discourse suggests that the facilitative effect of context observed in previous studies may help children resolve the *choice* between literal and metaphorical interpretations, rather than explain their general ability to derive metaphorical meaning. We therefore suggest that considering participants' responses to literal competitors, even if provided with disambiguating context, that literally biased responding should not be taken simply as evidence of a deficit, or even fragility, with metaphor. Instead, equal attention should be given to why literal meaning might be particularly salient or prioritized by children.

Developmental Trajectories

We replicated Pouscoulous and Tomasello's (2020) finding that 3-year-olds can understand attributional metaphors and extended it to more abstract metaphors such as functional and psychological ones. Previous research found that younger children performed worse on abstract metaphors compared to older children (Asch & Nerlove, 1960; Gentner, 1977; Winner et al., 1976), so when extending our paradigm, we included older children to be able to detect any such developmental differences. Though broadly, children of all ages and adults performed similarly, we did find some evidence to suggest that 3-year-olds less successfully derived metaphorical interpretations compared to 5- and 7-year-olds. However, we also found that 3-year-olds were less successful in the control condition (Experiment 3), so it could be that word knowledge instead of the ability to derive metaphorical interpretations explains these differences (see also Keil, 1986; Norbury, 2005; Vicente & Martín-González, 2021). Whether these small differences in our data resulted from simple word knowledge differences or differences in metaphor processing would require further research. However, for the purpose of our study, the fact that children performed above chance regardless of these potential differences—and that younger children were not more literally biased in the second experiment—highlights children's general ability with novel, nominal metaphors and suggests that literalist responses should not be taken as direct evidence of deficient metaphor comprehension.

Children of all ages and adults similarly passed our metaphorical selection task in Experiment 1 and chose literally in Experiments 2 and 3. However, even though surface accuracy levels appear comparable, the reaction time data indicate different processing strategies. Across all ages, participants responded more slowly in the metaphor versus distractor condition (i.e., Experiment 1) than in the literal versus distractor condition (i.e., Experiment 3). This finding aligns with research suggesting that deriving metaphorical meaning is more effortful (e.g., Noveck et al., 2001). We also found that reaction times overall decreased with age, which could be due to factors other than children's developing competence with metaphor comprehension. For instance, it might be that access and retrieval of conventional senses are generally slower in children than in adults.

To explore whether participants were still sensitive to the metaphorical meaning when literal options were available, we conducted a third experiment to see if children's literal selections were processed faster in this unambiguous context than in Experiment 2. We found that 3- and 5-year-olds, as well as adults, took longer to respond literally in Experiment 2 compared to Experiment 3. The fact that participants were highly literal in both studies but took significantly longer when the context was more ambiguous demonstrates their lingering sensitivity to the competing metaphorical alternatives despite their literalist responses.

However, we do not claim that this sensitivity can be extrapolated to reflect participants' full comprehension of the situation's ambiguity, as we might expect that to result in longer reaction times overall (as seemingly if one were sensitive to the ambiguity of the choice, then making that choice should take longer compared to situations without any ambiguity). Instead of suggesting that younger children's choices are literal because they cannot access metaphorical meaning, some of the developmental differences seen across the literature may result from developing sensitivities to referential ambiguity (see Long et al., 2021 for a similar discussion with their data with 13-year-olds). It could also be that this sensitivity interacts with the ability to overcome the less effortful choice (i.e., literal interpretation). Because we did not provide disambiguating discourse context we cannot speak to a clear distinction between these alternatives, especially in light of adults' literal responses in the ambiguous context. However, we still argue that decoupling these alternatives from children's metaphor comprehension provides a better explanation for the data. Even considering the curious reaction time shift in the 7-year-olds, where they appeared to choose literally in Experiment 2 just as quickly as in Experiment 3: It seems more likely that these older children were more influenced by the presence of a literal alternative than a loss of sensitivity to metaphorical meaning (see Lee et al., 2022 for similar findings with surface-meaning biases in 8-year-olds and Köder & Falkum, 2020 for accounts on related U-shaped developments in nonliteral language contexts).

Conclusion

Our study challenges the common assumption that literal responses represent fragile metaphor comprehension. Contrary to deficit-based expectations, young children successfully derived metaphorical interpretations, and we found no age-related decline in literalist responses within ambiguous contexts where both metaphorical and literal options were available. We cannot argue whether these findings generalize to all metaphor comprehension because we did not probe children's understanding of metaphor forms. However, given that even adults chose mostly literally when provided literal options, it is clear that literal options introduce a meaningful conflict that would likely persist across tasks. Consequently, previous research, which often included literal competitors, may have underestimated children's early metaphor comprehension abilities.

Ultimately, our findings call for a decoupling of literalist responses from metaphor comprehension ability and for researchers to more clearly distinguish between studies assessing sensitivity to metaphorical meaning from those investigating when metaphorical interpretations "ought" to be prioritized over literal ones. Further, we urge researchers who use literal options to interpret any literalist responding cautiously rather than treating it as evidence of figurative language difficulties, especially in

light of increasing evidence of children's early sensitivity to metaphorical meaning and harmful implications resulting from deficit-worded accounts.

References

Allison, P. (2012, September 10). When can you safely ignore multicollinearity? Statistical Horizons. https://statisticalhorizons.com/multicollinearity/

Almohammadi, A., Gaskins, D. K., & Rundblad, G. (2025). Metaphor comprehension in the acquisition of Arabic. *Journal of Child Language, 52*(2), 334-354. https://doi.org/10.1017/S0305000923000740

Asch, S. E., & Nerlove, H. (1960). The development of double function terms in children. In B. Kaplan & S. Wapner (Eds.), *Perspectives in Psychological Theory: Essays in Honor of Heinz Werner* (pp. 47-60). International Universities Press.

Barr, D. J. (2013). Random effects structure for testing interactions in linear mixed-effects models. *Frontiers in Psychology, 4,* 328. https://doi.org/10.3389/fpsyg.2013.00328

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

Billow, R. M. (1981). Observing spontaneous metaphor in children. *Journal of Experimental Child Psychology, 31*(3), 430-445. https://doi.org/10.1016/0022-0965(81)90028-X

Bloom, P. (2001). Precis of how children learn the meanings of words. *Behavioral and Brain Sciences, 24*(6), 1095-1103. https://doi.org/10.1017/S0140525X01000139

Box, G. E., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B, 26*(2), 211-243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ, 8,* e9414. https://doi.org/10.7717/peerj.9414

Bühler, D., Perovic, A., & Pouscoulous, N. (2018). Comprehension of novel

metaphor in young children with Developmental Language Disorder. *Autism & Developmental Language Impairments, 3*, 2396941518817229. https://doi.org/10.1177/2396941518817229

Cacciari, C., & Padovani, R. (2012). The development of figurative language. In M. Spivey, M. Joanisse, & K. McRae (Eds.), *The Cambridge handbook of psycholinguistics* (pp. 505–522). Cambridge University Press.

Callanan, M. A., Siegel, D. R., & Luce, M. R. (2007). Conventionality in family conversations about everyday objects. *New Directions for Child and Adolescent Development, 2007*(115), 83-97. https://doi.org/10.1002/cd.184

Clark, H. H. (1973). The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior, 12*, 135-142. https://doi.org/10.1016/S0022-5371(73)80014-3

Colston, H. L. (2020). Figurative language development/acquisition research: Status and ways forward. *Journal of Pragmatics, 156,* 176-190. https://doi.org/10.1016/j.pragma.2019.07.002

Davies, C., Lingwood, J., Ivanova, B., & Arunachalam, S. (2021). Three-year-olds' comprehension of contrastive and descriptive adjectives: Evidence for contrastive inference. *Cognition, 212*, 104707. https://doi.org/10.1016/j.cognition.2021.104707

Deák, G. O. (2006). Representing object functions: The cognitive basis of tool-use by children. In *Proceedings of the Fifth International Conference on Development and Learning, Bloomington, Indiana, USA*. http://quote.ucsd.edu/cogdevlab/files/2013/02/Deak_function_ICDL06.pdf

Deamer, F. M. (2013). *An investigation into the processes and mechanisms underlying the comprehension of metaphor and hyperbole* [Doctoral dissertation, University College London]. UCL Discover. https://discovery.ucl.ac.uk/id/eprint/1420534/

Di Paola, S., Domaneschi, F., & Pouscoulous, N. (2020). Metaphorical developing minds: The role of multiple factors in the development of metaphor comprehension. *Journal of Pragmatics, 156*, 235-251. https://doi.org/10.1016/j.pragma.2019.08.008

Eriksen, C. W. (1995). The flankers task and response competition: A useful tool for investigating a variety of cognitive problems. *Visual Cognition, 2*(2-3), 101-118.

https://doi.org/10.1080/13506289508401726

Estes, D., Wellman, H. M., & Woolley, J. D. (1989). Children's understanding of mental phenomena. *Advances in Child Development and Behavior, 22*, 41-87. https://doi.org/10.1016/S0065-2407(08)60412-7

Faitaki, F., & Murphy, V. (2019). Oral language elicitation tasks in applied linguistics research. *The Routledge Handbook of Research Methods and Applied Linguistics (1st ed.)*. https://doi.org/10.4324/9780367824471-31

Field, A. (2005). *Discovering statistics using SPSS (2nd ed.)*. Sage Publications, Inc.

Fox, J., & Monette, G. (1992). Generalized collinearity diagnostics. *Journal of the American Statistical Association, 87*(417), 178-183. https://doi.org/10.1080/01621459.1992.10475190

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The Wordbank project.* MIT Press. https://doi.org/10.7551/mitpress/11577.001.0001

Gardner, H., Kircher, M., Winner, E., & Perkins, D. (1975). Children's metaphoric productions and preferences. *Journal of Child Language, 2*(1), 125-141. https://doi.org/10.1017/S0305000900000921

Gardner, H., & Winner, E. (1978). The development of metaphoric competence: Implications for humanistic disciplines. *Critical Inquiry, 4*(1), 123-141. https://doi.org/10.1086/447976

Gentner, D. (1977). If a tree had a knee, where would it be? Children's performance on simple spatial metaphors. *Papers and Reports on Child Language Development, 13,* 157-164.

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: performance of children 3 1/2-7 years old on a Stroop-like day-night test. *Cognition, 53*(2), 129-153. https://doi.org/10.1016/0010-0277(94)90068-X

Gibbs, R. W., & Colston, H. L. (2012). *Interpreting figurative meaning.* Cambridge University Press. https://doi.org/10.1017/CBO9781139168779

Gottfried, G. M. (1997). Using metaphors as modifiers: Children's production of

metaphoric compounds. *Journal of Child Language, 24*(3), 567-601. https://doi.org/10.1017/S0305000997003176

Grice, H. P. (1975). Logic and conversation. In P. Cole & J. L. Morgan (Eds.), *Syntax and semantics: Speech acts* (pp. 41-58). Brill.

Harris, P. L., de Rosnay, M., & Pons, F. (2005). Language and children's understanding of mental states. *Current Directions in Psychological Science, 14*(2), 69-73. https://doi.org/10.1111/j.0963-7214.2005.00337.x

Harrison, X. A., Donaldson, L., Correa-Cano, M. E., Evans, J., Fisher, D. N., Goodwin, C. E. D., Robinson, B. S., Hodgson, D. J., & Inger, R. (2018). A brief introduction to mixed effects modelling and multi-model inference in ecology. *PeerJ 6*:e4794. https://doi.org/10.7717/peerj.4794

Horowitz, A. C., Schneider, R. M., & Frank, M. C. (2018). The trouble with quantifiers: Exploring children's deficits in scalar implicature. *Child Development, 89*(5), e572-e593. https://doi.org/10.1111/cdev.13014

Keil, F. C. (1986). Conceptual domains and the acquisition of metaphor. *Cognitive Development, 1*(1), 73-96. https://doi.org/10.1016/S0885-2014(86)80024-7

Köder, F., & Falkum, I. (2020). Children's metonymy comprehension: Evidence from eye-tracking and picture selection. *Journal of Pragmatics, 156*, 1-13. https://doi.org/10.1016/j.pragma.2019.07.007

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest package: Tests in linear mixed effects models. *Journal of Statistical Software, 82*(13), 1-26. https://doi.org/10.18637/jss.v082.i13

Lee, R., Chambers, C. G., Huettig, F., & Ganea, P. A. (2022). Children's and adults' use of fictional discourse and semantic knowledge for prediction in language processing. *PLOS One, 17*(4), e0267297. https://doi.org/10.1371/journal.pone.0267297

Levorato, M. C., & Cacciari, C. (2002). The creation of new figurative expressions: Psycholinguistic evidence in Italian children, adolescents and adults. *Journal of Child Language, 29*(1), 127-150. https://doi.org/10.1017/S0305000901004950

Lind, M., Simonsen, H. G., Hansen, P., Holm, E., & Mevik, B.-H. (2013). "Ordforrådet"-En leksikalsk database over et utvalg norske ord ("Ordforrådet"-A lexical

database of a selection of Norwegian words). *Norsk Tidsskrift for Logopedi, 59*(1), 18-26.

Lo, S., & Andrews, S. (2015). To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology, 6*, 1171. https://doi.org/10.3389/fpsyg.2015.01171

Long, M., Shukla, V., & Rubio-Fernandez, P. (2021). The development of simile comprehension: From similarity to scalar implicature. *Child Development, 92*(2), e1020-e1035. https://doi.org/10.1111/cdev.13507

Løvstakken, E., & Neff, M. B. (2024, May 1). Novel metaphor stimuli for children. OFSHome. https://doi.org/10.17605/OSF.IO/VJN42

The ManyBabies Consortium. (2020). Quantifying sources of variability in infancy research using the infant-directed-speech preference. *Advances in Methods and Practices in Psychological Science, 3*(1), 24-52. https://doi.org/10.1177/2515245919900809

Matthews, D. (2014). *Pragmatic development in first language acquisition (Vol. 10)*. John Benjamins Publishing Company.

Martín-González, I., Ronderos, C. R., Castroviejo, E., Schroeder, K., Lossius-Falkum, I., & Vicente, A. (2025). That kid is a grasshopper! Metaphor development from 3 to 9 years of age. *Journal of Child Language*, *52*(4), 945-970. https://doi.org/10.1017/S0305000924000187

Norbury, C. F. (2005). The relationship between theory of mind and metaphor: Evidence from children with language impairment and autistic spectrum disorder. *British Journal of Developmental Psychology, 23*(3), 383-399. https://doi.org/10.1348/026151005X26732

Noveck, I. A., Bianco, M., & Castry, A. (2001). The costs and benefits of metaphor. *Metaphor and Symbol, 16*(1-2), 109-121. https://doi.org/10.1080/10926488.2001.9678889

Özçalişkan, Ş. (2005). On learning to draw the distinction between physical and metaphorical motion: Is metaphor an early emerging cognitive and linguistic capacity? *Journal of Child Language, 32*(2), 291-318. https://doi.org/10.1017/S0305000905006884

Peirce, J., Gray, J. R., Simpson, S., MacAskill, M., Höchenberger, R., Sogo, H., Kastman, E., & Lindeløv, J. K. (2019). PsychoPy2: Experiments in behavior made easy. *Behavior Re-search Methods, 51*(1), 195-203. https://doi.org/10.3758/s13428-018-01193-y

Pouscoulous, N. (2011). Metaphor: For adults only? *Belgian Journal of Linguistics, 25*(1), 51-79. https://doi.org/10.1075/bjl.25.04pou

Pouscoulous, N., & Tomasello, M. (2020). Early birds: Metaphor understanding in 3-year-olds. *Journal of Pragmatics, 156*, 160-167. https://doi.org/10.1016/j.pragma.2019.05.021

Quinn, G. P., & Keough, M. J. (2002). *Experimental designs and data analysis for biologists.* Cambridge University Press. https://doi.org/10.1017/CBO9780511806384

R Core Team (2021). *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Ronderos, C. R., Falkum, I. L., Noveck, I. A., & Mathisen, E. (2022, June 3). Development of imprecise usages of absolute adjectives. *OSF Registries.* https://doi.org/10.17605/OSF.IO/KUGPF

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: overconfident estimates in mixed models. *Behavioral Ecology, 20*(2), 416-420.

Shanks, D. R. (2010). Learning: From association to cognition. *Annual Review of Psychology, 61*, 273-301. https://doi.org/10.1146/annurev.psych.093008.100519

Simonsen, H. G., Kristoffersen, K. E., Bleses, D., Wehberg, S., & Jørgensen, R. N. (2014). The Norwegian Communicative Development Inventories: Reliability, main developmental trends and gender differences. *First Language, 34*(1), 3-23. https://doi.org/10.1177/0142723713510997

Skordos, D., & Papafragou, A. (2016). Children's derivation of scalar implicatures: Alternatives and relevance. *Cognition, 153*, 6-18. https://doi.org/10.1016/j.cognition.2016.04.006

Sperber, D., & Wilson, D. (2008). A deflationary account of metaphors. In R. W. Gibbs Jr. (Ed.), *The Cambridge Handbook of Metaphor and Thought* (pp. 84-105).

Cambridge University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

Van Herwegen, J., Dimitriou, D., & Rundblad, G. (2013). Development of novel metaphor and metonymy comprehension in typically developing children and Williams syndrome. *Research in Developmental Disabilities, 34*(4), 1300-1311. https://doi.org/10.1016/j.ridd.2013.01.017

Vicente, A., & Falkum, I. L. (2021). Accounting for the preference for literal meanings in ASC. *Mind & Language, 38*(1), 119-140. https://doi.org/10.1111/mila.12371

Vicente, A., & Martín-González, I. (2021). El sesgo literalista en las condiciones del espectro autista: revisión de las teorías existentes [The literalist bias in the autistic spectrum conditions: Review of existing account]. *Studies in Psychology, 42*(4), 298-333. https://doi.org/10.1080/02109395.2021.1909248

Vosniadou, S. (1987). Children and metaphors. *Child Development, 58*(3), 870-885. https://doi.org/10.2307/1130223

Vosniadou, S. (1989). Context and the development of metaphor comprehension. *Metaphor and Symbol, 4*(3), 159-171. https://doi.org/10.1207/s15327868ms0403_4

Vosniadou, S., Ortony, A., Reynolds, R. E., & Wilson, P. T. (1984). Sources of difficulty in the young child's understanding of metaphorical language. *Child Development, 55*(4), 1588-1606. https://doi.org/10.2307/1130028

Wilson, D., & Carston, R. (2007). A unitary approach to lexical pragmatics: Relevance, inference and ad hoc concepts. In Noel Burton-Roberts (Ed.), *Pragmatics* (pp. 230-259). Palgrave-Macmillan.

Wilson, D., & Sperber, D. (2006). Relevance theory. In K. Brown (Ed.), *The handbook of pragmatics* (pp. 607-632). Blackwell Publishing.

Winner, E. (1997). *The point of words: Children's understanding of metaphor and irony*. Harvard University Press. (Original work published 1988)

Winner, E., Engel, M., & Gardner, H. (1980). Misunderstanding metaphor: What's the problem? *Journal of Experimental Child Psychology, 30*(1), 22-32.

https://doi.org/10.1016/0022-0965(80)90072-7

Winner, E., Rosenstiel, A. K., & Gardner, H. (1976). The development of metaphoric understanding. *Developmental Psychology, 12*(4), 289–297. https://doi.org/10.1037/0012-1649.12.4.289

Zufferey, S. (2015). *Acquiring pragmatics: Social and cognitive perspectives.* Routledge.

Data, Code and Materials Availability Statement

All data, materials, and analysis scripts are available on the Open Science Framework at https://osf.io/jkq9w/, except for children's demographic profiles. These are instead provided as aggregated summaries in the Supplementary Information section of the manuscript. As these data were considered non-anonymized personal data, they could not be shared in raw form.

Ethics Statement

All experimental protocols adhered to The National Committee for Research Ethics in the Social Sciences and the Humanities guidelines and have been approved by the Norwegian Agency for Shared Services in Education and Research (Reference Number: 596365). All participants gave informed written consent before taking part in the study.

Authorship and Contributorship Statement

Mary Beth Neff: conceptualization, methodology, software, validation, resources, writing – original draft, writing – review & editing, visualization, project administration, data curation, investigation, and formal analysis. **Ingrid Lossius Falkum**: supervision, methodology, resources, writing – review & editing, funding acquisition.

Funding

This research was supported by the European Union's Horizon 2020 research and innovation program, under grant agreement 853211 (an ERC Starting Grant 2019 awarded to Ingrid Lossius Falkum).

Acknowledgements

We would like to thank Roger Mundry for advice on statistical simulations, as well as Nausicaa Pouscoulous for her help with replicating the original study. We would also like to thank all of the participating families and the DEVCOM research assistants who made this research possible. In particular, we thank Eline Løvstakken for her help in creating the metaphorical stimuli, Emma Krane Mathisen and Henriette Johansen for voicing the characters, Tyler Green and Peter Philipp for trouble-shooting PsychoPy, and Håkon Bjørn Due for testing all those participants!

Supplementary Information

Table S1. ***Children's average language development demographics per experiment and age group***

age	exp	*N*	% female	% exposure to lang1	% lang1 = norsk	% w/ language dev't issues	% w/ dev't issues	% care-giver A w/ master's edu	% care-giver B w/ master's edu
3	3	16	19	95	100	19	0	50	44
3	2	16	38	97	94	0	0	50	25
3	1	18	50	84	92	11	0	50	56
5	3	17	29	84	91	6	6	31	44
5	2	17	24	83	94	12	0	29	53
5	1	13	54	82	92	15	0	33	33
7	3	14	21	90	96	14	7	15	31
7	2	24	50	83	94	13	0	50	32
7	1	10	70	96	85	40	10	14	0

Note. All children were born in Norway. Sample sizes will vary from the original experimental groups as the form was optional for caregivers and could include excluded participants (for practice failures). We did not collect these data from adult participants.

Table S2. ***Average inhibitory control task scores per experiment / age group***

Age Group	Experiment	*N*	*Mean*	*SD*
Adult	Experiment 1	26	0.81	0.39
	Experiment 2	26	0.88	0.32
	Experiment 3	16	0.83	0.38
7	Experiment 1	25	0.97	0.06
	Experiment 2	26	0.98	0.05
	Experiment 3	15	0.96	0.07
5	Experiment 1	27	0.74	0.31
	Experiment 2	26	0.74	0.32
	Experiment 3	16	0.83	0.15
3	Experiment 1	28	0.76	0.23
	Experiment 2	28	0.67	0.33
	Experiment 3	17	0.54	0.32

Note. Sample sizes will vary from the original experimental groups as this task was always run last, some participants withdrew before taking the task, and data is from all possible subjects (including replacements and those excluded for practice failures). Task = DayNight for children and = Flanker for adults.

Table S3. ***Proportion of metaphorical picture selections in Experiment 1 by item and metaphor types for 3-year-olds***

Age	Metaphor Type	Metaphor Item Statement in English	Mean Correct (proportion)
3	Replicated Attributional	The tower with the hat	1.00
		The car with the backpack	0.73
		The carrot with the hair	1.00
		The dog with the brown shoes	0.92
		The bottle with the belly	0.85
	New Attributional	The tree with the arms	1.00
		The head with the spaghetti	0.89
		The sky with the cookie	0.54
		The hot cocoa with the pillows	0.85
		The tree with the fried eggs	0.42
	Functional	The frog with the umbrella	0.92
		The fox with the flashlight	0.88
		The monkey with the hammer	0.69
		The squirrel with the cup	0.77
		The bug on the plane	0.96
	Psychological	The plant that is sad	0.73
		The balls that are happy	0.65
		The drawing that is angry	0.81
		The bag that is sleeping	0.69
		The boy that is on fire	0.81

Table S4. ***Proportion of metaphorical picture selections in Experiment 1 by item and metaphor types for 5-year-olds***

Age	Metaphor Type	Metaphor Item Statement in English	Mean Correct (proportion)
5	Replicated Attribu-tional	The tower with the hat	1.00
		The car with the backpack	0.93
		The carrot with the hair	1.00
		The dog with the brown shoes	1.00
		The bottle with the belly	1.00
	New Attributional	The tree with the arms	0.89
		The head with the spaghetti	1.00
		The sky with the cookie	0.89
		The hot cocoa with the pillows	1.00
		The tree with the fried eggs	0.07
	Functional	The frog with the umbrella	1.00
		The fox with the flashlight	0.85
		The monkey with the hammer	0.89
		The squirrel with the cup	0.93
		The bug on the plane	0.96
	Psychological	The plant that is sad	0.96
		The balls that are happy	0.93
		The drawing that is angry	0.96
		The bag that is sleeping	1.00
		The boy that is on fire	0.96

Table S5. ***Proportion of metaphorical picture selections in Experiment 1 by item and metaphor types for 5-year-olds***

Age	Metaphor Type	Metaphor Item Statement in English	Mean Correct (proportion)
7	Replicated Attribu-tional	The tower with the hat	1.00
		The car with the backpack	0.92
		The carrot with the hair	1.00
		The dog with the brown shoes	1.00
		The bottle with the belly	1.00
	New Attributional	The tree with the arms	1.00
		The head with the spaghetti	1.00
		The sky with the cookie	0.92
		The hot cocoa with the pillows	1.00
		The tree with the fried eggs	0.28
	Functional	The frog with the umbrella	1.00
		The fox with the flashlight	1.00
		The monkey with the hammer	0.92
		The squirrel with the cup	1.00
		The bug on the plane	0.96
	Psychological	The plant that is sad	1.00
		The balls that are happy	0.88
		The drawing that is angry	1.00
		The bag that is sleeping	1.00
		The boy that is on fire	0.92

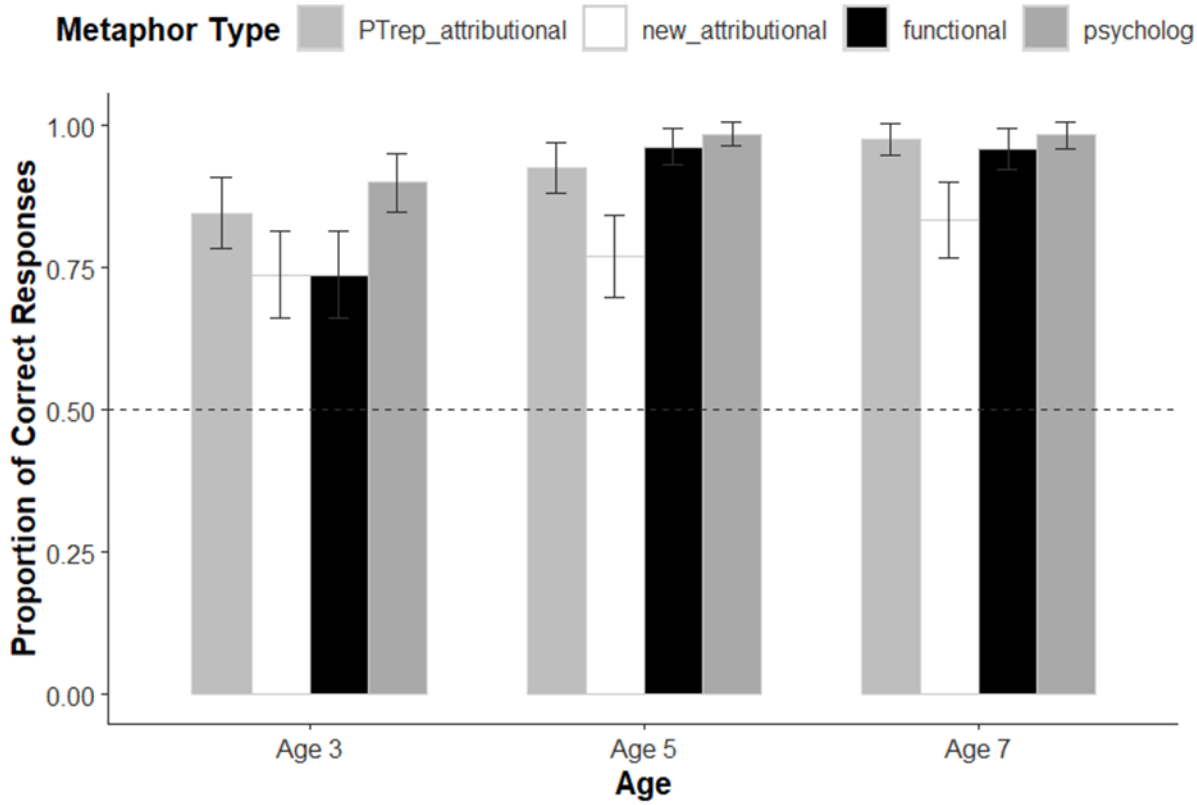

Figure S1. ***Proportion of metaphorical picture selections in Experiment 1 across all metaphor types and ages.***

License