How do theory of mind and language skills impact metaphoric reference comprehension during children's school-age years

Nicolas PETIT[1,2,*], Valentina BAMBINI[3], Luca BISCHETTI[3],
Jérôme PRADO[2] and Ira NOVECK[4]

[1] *Hospital Le Vinatier (Bron, FRANCE)*
[2] *Centre de Recherche en Neurosciences de Lyon (Lyon, FRANCE)*
[3] *Laboratory of Neurolinguistics and Experimental Pragmatics (NEP), University School for Advanced Studies IUSS (Pavia, ITALY)*
[4] *Laboratoire de Linguistique Formelle (Paris, FRANCE)*

* Corresponding author: npetit.ortho@gmail.com

Authors manuscript (July 9, 2023). This paper has not been peer-rewiewed.

Metaphor comprehension development, a long-standing topic in the cognitive science, is the source of multiple strands of research. These provided evidence both for the presence of early precursors in young children and for a progressive and relatively late mastery of metaphor, which this work focuses on. Moreover, research in the field of atypical development has put to the fore two important cognitive bases for metaphor comprehension, that are theory of mind (TOM) and structural language, but these have remained relatively understudied in typical development, and not yet in a joint perspective. To address this issue, this work introduces a new tablet-based, orally presented metaphoric reference task, which allows for response times collection. In Experiment 1, we confirmed the viability and developmental sensitivity of this task in typically developing 7- and 10-year-old children (N=89). In Experiment 2, we used a version of this task with 248 typically developing 6- to 11-year-olds, together with language and non-verbal TOM tasks. Results showed a non-linear developmental trend in metaphor comprehension, with a focused improvement between 7 and 9 years of age. Mixed-effect models explaining children's accuracy and response times revealed a critical developmental shift with respect to the cognitive predictors of metaphor comprehension. The data showed that TOM is a reliable predictor of successful performance among younger children but fades away, at which point language becomes a predictor of success on the task among older children. This study underlines the importance of considering developmental perspectives when referring to the cognitive bases of metaphor skills.

Keywords: metaphor; experimental pragmatics; pragmatic development; theory of mind; reference; language development

Public significance statement: This study deepens our understanding of school-age as a key period for the development of metaphors, by evidencing a spurt in metaphors comprehension between the ages of 6 and 11, together with a shift in its associations with other cognitive skills. While the ability to read one's mind is the strongest predictor of metaphor comprehension in younger children, this influence appears gradually replaced by that of formal language skills.

Research on the development of metaphor comprehension can be reduced to three overlapping, but largely-independent, strands of investigations. Historically speaking, the first of the current strands concerns the development of metaphor comprehension among school-aged children. This line of work began with Piaget (1964) who claimed – based on findings with out-of-context proverbs (which are often but not always metaphorical) – that children need to reach adolescence before they could understand and produce metaphoric utterances. Piaget's account would eventually spur many researchers to report observations – often from their own children – contradicting his claim (see e.g., Gentner, 1988) and it prompted many others to carry out developmental studies showing that young children are indeed much more competent with metaphor than he proposed (e.g., Vosniadou, 1987). As research studies accrued, the findings have eventually revealed support for two complementary claims: one is that metaphor comprehension indeed continues to improve with age, to the point that children become adult-like progressively. However, research also shows that there is some rudimentary ability among children who are as young as 3 years of age to use and understand metaphor (Pouscoulous, 2011; Pouscoulous & Tomasello, 2020). This accounts for the first two of the strands.

Along the way, the *nature* of metaphor itself became central to investigations. That is, researchers took an interest in children's metaphor performance in order to a) investigate the extent to which the properties of metaphor could account for children's difficulties and to; b) better understand the underlying abilities that facilitate metaphor comprehension. This is what ultimately led to the third strand. To make this clear, consider two examples that underline the importance of psychological factors in metaphor comprehension. The first comes from Winner et al. (1976) who showed how psychological metaphors, e.g., *The prison guard had become a hard rock*, are more difficult for children to understand than physical metaphors, e.g., *hair is spaghetti* (for more recent results in this vein, see Lecce et al., 2019). The second, and arguably more impactful, example in this line comes from developmental research showing that figurative language (such as metaphor) could be impaired in autistic compared to neurotypical participants (Happé, 1993; for reviews see Kalandadze et al., 2018 or Morsanyi et al., 2020). The proposal motivating Happé's work was that theory of mind (TOM, the ability to understand an interlocutor's intended meaning) is a critical factor for metaphor comprehension. Once researchers considered a role for theory of mind in metaphor comprehension, it became only natural to test this claim against alternative factors that could account for developmental changes, such as structural language abilities (Norbury, 2005).

By viewing the metaphor development literature in this three-pronged way, it becomes clearer why it is populated with such a wide variety of studies. Moreover, by appreciating that there are multiple strands, one can better appreciate how the literature has evolved. Given the breadth of investigations on metaphor development, it would obviously be difficult for any single study to address all three strands. It is nonetheless worthwhile to attempt to interlace, at the very least, two of them. Here, we introduce a task that investigates the development of metaphoric (versus synonymic) referential comprehension during school age, which is tested through participants' choice of picture in the wake of a referential expression, and we go on to consider the relative influence of two underlying factors that are often assumed to be critical to metaphor: theory of mind and structural language comprehension.

We break down the remainder of this Introduction into three parts. First, we summarize studies that investigate school-aged children with tasks that endeavor to capture developmental progress with metaphor. Second, we briefly review findings that aim to uncover the substrates of metaphor comprehension. This will lead to a discussion concerning the way theory of mind as well as language abilities affect participants' growing metaphor comprehension. Finally, we will introduce our two Experiments, which essentially investigate children's accuracy in identifying a metaphoric reference while listening to orally-presented vignettes via a tablet.

The development of metaphoric abilities among school-aged children

A large number of studies now assume that school-aged children do in fact progressively gain competence with metaphor with age. For the purposes of providing background on our study, we describe two streams of research that have been particularly helpful in providing insights into metaphor development. The first relies on a paradigm that asks children to *detect* and metalinguistically *explain* metaphoric relations among triads. Kogan et al. (1980) introduced what they called the *Metaphor Triad Task* (the MTT), in which young participants are presented with three items that allow for up to three pairing possibilities, one of which is metaphoric in character while the remaining two reflect other sorts of relationships. In this task, participants are requested, by an experimenter, to select the best possible pairing and to explain it. They are also asked to continue (and are thus slightly prodded) so that they ultimately always consider the metaphoric one. Consider the triad *fish, winding river,* and *snake*. Here, the latter two would be expected to be the source of metaphoric pair. This technique captured children's growing metaphoric abilities reliably whether the items in the triad were presented pictorially or with words (Kogan & Chadrow, 1986).

To appreciate this technique in a more recent study, consider work from Willinger's group (Deckert et al., 2019; Willinger et al., 2019). Inspired by the prior work, Willinger et al. (2019) tested 164 seven-, nine- and eleven-year-olds with a form of the MTT and recorded three different scores for each triad: whether the metaphorical relation was identified or not (*identification*), whether it was correctly explained or not (*explanation*), and whether it was preferred over other associations, i.e., whether it was identified first (*preference*). Their findings revealed that the eleven-year-olds had higher metaphor comprehension quality and preference scores than the younger groups. Deckert et al. (2019) tested 298 typically developing children, between the ages of 7 and 11, and observed different turning points, depending on the measure they considered. A first turning point was identified at 8.2 years of age, after which children *identified* the metaphoric pairs and *explained* them better, and a second was observed at 10.2 years old, after which they increasingly *preferred* metaphoric pairs over non-metaphoric ones. This resonated with Willinger and colleagues' (2019) observations. The authors explain that these gaps might correspond to changes in the *nature* of the processing strategies of metaphors – and not to a continuous development – that they link to brain structure modifications which occur over the course of childhood. Willinger et al. (2019) also linked this change to the concurrent increase of the communicative experience, at an age at which relationships are known to become more complex.

A second stream of research that investigates school-aged children's increasing abilities with metaphor comes from studies that use vignettes that ultimately include metaphoric lines of text that refer back to a previously introduced element, a technique that was initially introduced with adults by Gibbs's (1990). For example, in one vignette, Gibbs (1990) used the word *creampuff* in the sentence *The creampuff didn't even show up* to refer metaphorically to a previously mentioned boxer. Noveck et al. (2001) employed this sort of referential paradigm to investigate the development of metaphor comprehension. In their first experiment, Noveck et al. (2001) presented 8- to 12-year-old French-speaking children with written vignettes whose penultimate sentence included a reference to a previously mentioned element. Consider one of their stories about school pupils in a pool who are later referred to as either "toads" metaphorically or else as "children" synonymically. The authors then determined, through yes/no questions, whether participants understood the reference (e.g., "was it the pupils who were sent to the side of the pool?"). They observed that, while referential abilities improved with age in general, metaphoric references consistently prompted more errors than the synonymic ones, until around 12 years old, where the gap appears to close. In a second experiment, they implemented this paradigm in the form of a self-paced computer task, in order to capture reading times for the referential sentences. Their data revealed – for children as well as adults – that metaphoric references were associated with longer reading times than their synonymic controls; this indicates that there is a cost in making metaphoric references. Similar referential tasks have been used in other developmental studies (Seigneuric et al., 2016; Tonini et al., 2023) and led to similar studies with adults that confirmed a temporal cost associated with metaphor processing (Almor et al., 2007; Heredia & Cieślicka, 2015; though see Carston & Yan, 2022).

Van Herwegen et al. (2013) designed a similar reference task in which metaphors (as well as metonymies) were ultimately embedded in orally presented stories that were six or seven sentences in length and that were accompanied by drawings along the way. At the end of each story, a question was presented to children about the meaning of a metaphor used in the story. The participants were presented with three new drawings and answered by selecting one: one picture represented the intended object of the metaphor (e.g., a car which was described as a *turtle*), one represented a literal interpretation of the metaphorical expression (a turtle), and the third was a distractor. The 3 to 17 years-old participants' picture-choice results showed that metaphor comprehension increased significantly with age.

Taken together, these studies show that children's competencies with metaphor grow over time. Whether it concerns justifying the pairing of words or the comprehension of stories, metaphoric competence appears to increase during children's school age years. Another question then becomes what motors participants' progress?

On the substrates of metaphor comprehension among children

Here, we turn to the variability of children's performance across studies as a function of task properties (for a recent review, see Kalandadze et al., 2019). As we indicated above, one of the earliest proposals concerning the predictors of children's metaphor comprehension was through the prism of autism, which is known to be associated with a more literal understanding of language. This was what led Happé (1993) to link autistic individuals' difficulties in metaphor comprehension to their presumed impairment with respect to theory of mind. Indeed, in line with Relevance Theory (Sperber & Wilson, 1996), accessing the meaning of a metaphor is achieved through inferences about the speaker's intention.

One recent study that investigates the role of TOM in metaphor comprehension development comes from Lecce et al. (2019), who introduced a new Physical and Mental Metaphor (PMM) task while using metaphors of the form "X is Y." Critical to the study was a comparison of *physical* metaphors, which rely on features such as appearance or behavior (e.g., *dancers are butterflies*) to *mental* metaphors, which rely on psychological traits (e.g., *daddy is a volcano*). In their task, 217 9- to 12-year-old children were aurally presented with metaphors and asked to explain their meaning. Two different measures were coded from each response: whether the explanation was acceptable (*accuracy*) and whether the explanation relied on physical or mental features (*interpretation*). Results revealed that *accuracy* scores were high overall and, interestingly, that regarding *interpretation* scores, the nine-year-olds, the youngest group in the sample, performed at rates that were inferior to the other age-groups on mental, but not physical, metaphors. That is, while children provided plausible explanations overall, older children appear to rely on psychological features more than the youngest children, leading to a clear developmental effect between 9 and 11 years of age. Children from this study were also directly tested with the Strange Stories task (Happé, 1994), a now classic advanced TOM measure. Interestingly, they observed that this TOM measure predicted mental (but not physical) metaphor interpretation at 9 years of age, after controlling for vocabulary, SES and working memory, but they found no such associations for 10- and 11-year-olds. They thus introduced the idea that TOM's influence on metaphor interpretation could be short-lived as children grow (see also Tonini et al., 2023). Accordingly, in a short-term longitudinal study, the same team would go on to observe an association between metaphoric explanations and theory of mind growth among 9-year-olds (Del Sette et al., 2020), confirming the likely importance of TOM for children's metaphor comprehension at that age.

As we indicated above, Happé's seminal work led to an alternative proposal from Norbury (2005), who argued that formal language skills were an important basis for metaphor comprehension. According to this view, semantic ability would be a better predictor of children's metaphor comprehension, since the semantic traits of the vehicle (e.g., the metaphoric meaning of *shark* in *my lawyer is a shark*) that need to be connected to those of the topic (*lawyer* in this example) are often not the most salient traits; this implies that sophisticated semantic representations are called for when discerning metaphors. Although work from these two seminal streams (the one centered on TOM and the other centered on semantic meaning or language) prompted some criticism, these abilities have emerged as the two most dominant candidates for the cognitive scaffolding of metaphor comprehension in the field of autism (for a review see Kalandadze et al., 2018) as well as for neurotypical development.

While considering the relative importance of TOM versus language abilities, Whyte & Nelson (2015) tested a group of 69 typically developing children between the ages of 5 and 12, on non-literal language (which included not only metaphor, but other pragmatic phenomena such as sarcasm and indirect requests). They observed that both theory of mind, measured with the Reading the Mind in the Eyes test (Baron-Cohen et al., 2001) and language, as measured with tests of vocabulary and syntax, are correlated with increasing comprehension. Based on Whyte & Nelson, one can say that the two abilities – TOM and linguistic abilities – have a role to play in children's growing abilities with non-literal language, but it is not clear whether the two work in tandem or separately. Thus far, no research, as far as we know, has aimed to disentangle the two in a single developmental study.

That said, researchers do occasionally try to separate out the influence of key abilities on metaphor development. For example, the previously-mentioned Seigneuric et al. (2016) study used a written referential metaphor task in school-aged children to describe how "poor" comprehenders (defined as children with poor text comprehension but average or above average decoding abilities and vocabulary) had lower accuracy rates when identifying metaphoric referents, as opposed to literal referents. This led them to conclude that comprehension difficulties are due to something other than decoding and vocabulary. Similarly, in Deckert et al.'s (2019) metaphor triad task described earlier, the authors reported (page 347) that "for age groups seven and eight, [metaphor] comprehension was not significantly predicted by verbal intelligence and linguistic competence […] whereas in age-groups nine and ten comprehension [it] was significantly predicted by verbal intelligence." In other words, *verbal intelligence* accounted for increases in metaphor comprehension among 9- and 10-year-olds more than among 7- and 8-year-olds. Data like these underscore the possibility that metaphor comprehension does not occur steadily and that specific competencies could have a passing influence on the development of metaphor comprehension, i.e., TOM and language skills could impose themselves on metaphor comprehension at specific points in a child's cognitive life.

Goals of the current study

The current work has two general goals. One is to confirm that a new variations of a self-paced metaphor-reference comprehension task, as developed here, provides developmental outcomes that are consistent with prior studies (e.g., Noveck et al., 2001; Seigneuric et al., 2016; Van Herwegen et al., 2013). This is important because the new variations to be used here amounts to a novel task. In each task item, participants listen to a vignette that introduce a narrative element (e.g., *a crying baby*), that will be referred to in the target sentence, at the end of the vignette. Crucially, this reference can be metaphorical (say, '*the fountain*' is used to refer to the crying baby), or else synonymic, as a function of the vignette opening context ('*the fountain'* could also be used to refer to some

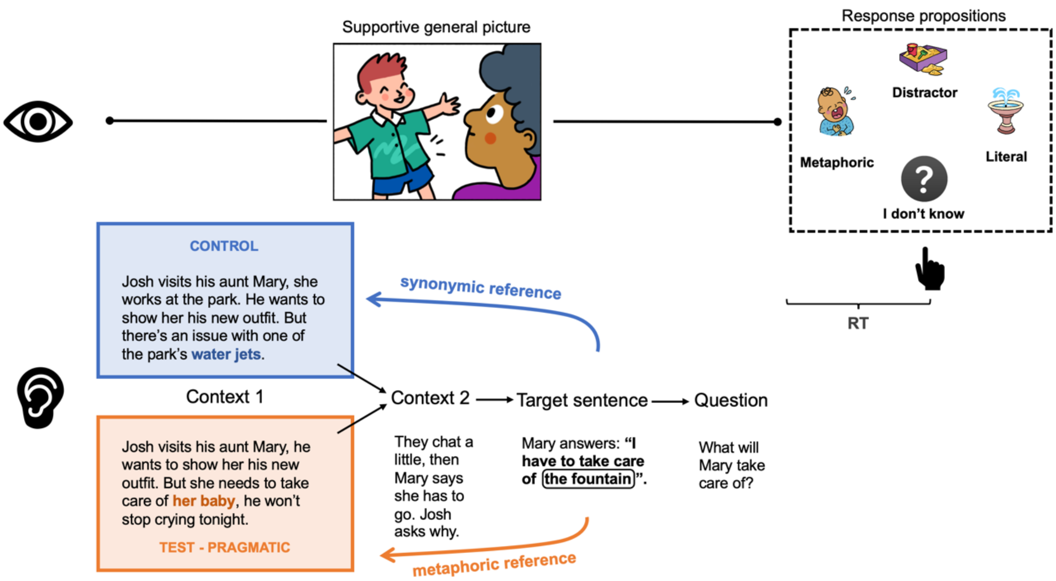


Figure 1: *Illustration of a vignette pair structure (translated from French to English) depicting what a participant sees (top) and hears (bottom)*

water jets that were introduced instead of a crying baby). Then, participants are asked a question to be answered through a picture selection, that can be timed and thus reveal how well and how fast they comprehend this reference. Importantly, in the metaphor condition, the intended target (a crying baby) is represented alongside a foil, which is a pictorial representation of the reference itself (a fountain, see **Figure 1**, to be detailed in the methods).

To test the viability and developmental sensitivity of the metaphor reference task, Experiment 1 tests two relatively distant age groups of 7- and 10-year-olds. We expect to find the older participants' performance with the task to reflect greater competence than the younger ones'. Likewise, we expect that the picture choice to be slower for correct metaphoric choices when compared to correct synonymic choices.

The other goal of the current study concerns deepening our knowledge about metaphor comprehension during the school-age years. This leads to Experiment 2, which investigates a larger sample of children (between the ages of 6 and 10) as they carry out the metaphor-reference task. Besides aiming to confirm the developmental findings from Experiment 1 in finer detail, it will also assess children's formal language skills and TOM, in order to gauge the contributions that these two factors make on metaphor comprehension. Previous studies focused on the importance of one or the other of those potential predictors. Consider Lecce et al., (2019) or Tonini et al. (2023), who use language measures as a control variable, but focus their investigation on the role of TOM. The roles of these two predictors have yet to be studied jointly in a single study.

To underline what else is novel in this work, we point out two other features of the task's implementation. One is that the task was prepared on a tablet which allows young participants to work at their own pace in the comfort of their classroom. The other is that, while the linguistic abilities of children in Experiment 2 will be assessed through a standard measure, the *BILO-3C Oral comprehension* task (Khomsi et al., 2007), the TOM measure will be determined by another tablet-based adaptation, the Langdon & Coltheart [1999]'s Picture Sequencing Task (PST), which was validated in [*Authors study, submitted, masked for review*]. As the title of the tablet-PST suggests, it requires participants to put pictures in a thematic order. One class of items in the tablet-PST requires participants to represent a false belief (e.g., a character who mistakenly assumes that their dog ate a pie that was cooling when in fact it was stolen by someone through the kitchen's window). Prior studies show that such false belief representations are harder to sequence than controls. Importantly, this task relies on minimally verbal material, which rules out the possibility of a language confound. Below, we describe the Experiments, as well as its adjacent tasks, in greater detail.

EXPERIMENT 1

The main goal of this Experiment is to test whether a novel variation of the metaphoric reference task provides developmental effects that are consistent with prior studies, in which children's metaphor comprehension increases with age. In the current Experiment, participants are aurally presented short vignettes that include a character and a set of events. At the end of the vignette, an earlier-mentioned element is referred to in a metaphoric or synonymic way. To test participants' comprehension, they are then presented with a picture-choice task that, for the metaphor condition, juxtaposes a picture showing the intended figurative referent with one showing a literal representation of the reference. Importantly, this task allows children to work independently via a tablet that collects accuracy data as well as latencies.

Method

Participants

Eighty-nine children took part in this study. All were native French speakers recruited from a local private school. None had neurodevelopmental disorders (learning, autism spectrum, attention-deficit, or language-related disorder) nor any sensory (visual or auditory) or motor disorder that would prevent them from using a tablet, as declared by parents. The study included two age groups: young children from 1st and 2nd grade (N = 48, 48% girls, mean age = 7;0, SD = 0;7) and older children from 4th and 5th grade (N = 41, 32% girls, mean age = 10;0, SD = 0;7). Ethnicity is not reported since the collection of such information is forbidden by

French laws. All children were free to participate, and their parents provided informed non-opposition, in accordance with local research ethics rules.

Materials

We developed a referential task inspired by Noveck et al.'s (2001) and Van Herwegen et al.'s (2013) paradigms. Each item consists of an orally-presented vignette in French, uttered by the first author and split into three parts (see **Figure 1**). In the first part of the vignette, two to three sentences provide background including mention of a noun phrase that will be referred back to in the last part of the vignette in the target sentence. It is this later reference that will be either metaphoric or synonymic.

Each vignette is introduced with two slightly different contexts. To put it another way, the target reference is designed to be either a metaphoric or synonymic reference with respect to an earlier mentioned element that depends on the initial context (e.g., the reference *fountain* in Figure 1 can be used to refer back to a *crying baby* in the metaphoric condition or to *water jets* in the synonymic condition). The metaphoric target expressions were designed to be novel. Note that, unlike in Noveck et al. (2001), which used different target words to refer to a singular previously-mentioned element, in the current task the *initial referents* vary across two slightly different contexts before setting up a singular target reference.

Another general feature of the paradigm includes the display of a picture that is thematically related to the vignette. This is done in order to engage the child's attention. Technically speaking, the task is self-paced; to advance the story, participants are required to press a speaker button in order to hear the next part of the story. Participants can listen to a segment as often as they pleased through a 'repeat' button.

After the vignette's presentation is completed, a question about the target reference is uttered by a female voice (for example, in **Figure 1**, she asks "*What will Mary take care of?*"). Immediately after the end of the question, four pictorial options are displayed on the screen. From the point of view of the metaphoric reference condition in **Figure 1**, the picture options reflect a) a representation of the metaphoric reference that harks back to the earlier referent (the crying baby), b) a representation of the literal meaning of the target reference (a picture of a fountain), c) a distractor that is distantly related to the story content (a sandbox which could be found in a park), and; d) an "I don't know" symbol. Both the participants' picture selection and their selection responses times are automatically recorded by the tablet.

For expository purposes, from here on when we refer to the two versions of each vignette, we will call them *metaphoric* and *synonymic* because the reference in the target sentence refers back to an earlier mentioned element in one of those two ways. When we consider the picture options that are available, we will refer to the intended meaning of the target reference. Thus, after a vignette makes a *metaphoric* reference, an accurate response ultimately involves choosing a picture that depicts the original referent (the *crying baby*). On the other hand, after a vignette makes a *synonymic* reference, an accurate response involves choosing a picture that corresponds to the literal meaning of the target reference. The upshot of this approach is that in the wake of a metaphoric reference, we expect the literal representation of the target reference to have some appeal for participants but at decreasing rates with age. In contrast, note that the correct response in the synonymic condition is expressed through a representation of the literal target reference only (the fountain). For this condition, both of the two remaining pictorial options are not related to the story.

Twelve vignettes were designed so that each could produce two versions, one that provides for a metaphoric reference and one a synonymic one. Efforts were made to limit the linguistic demands of the task, with simple sentences and high frequency words. Stories are matched (see **Table S1.1** in supplementary materials) for (1) length, (2) readability, (3) mean frequency of content words (nouns, verbs, adjectives, adverbs) in an oral French corpus, and (4) frequency of referents in school textbooks.

As far as what an individual participant sees, each vignette was presented only once (in one of its two versions). The experiment was designed to distribute equal amounts of metaphoric and synonymic versions; so, four lists of items were prepared. In each, six vignettes were ultimately presented in their metaphoric version and six in their synonymic one. All told, each vignette theme appeared in their metaphoric version in two lists and in their synonymic version in the two other lists, so that each version of each vignette appeared together with two different sets of stories. The presentation order of the 12 vignettes across the four lists (once they were assigned their metaphoric or synonymic status) was based on a randomized procedure, but to control for potential trial-*order* effects, each list order was also reversed, so that ultimately 8 different lists were distributed.

To familiarize participants with the task and response format, an additional training vignette, involving a synonymic reference, was added prior to the experimental session. This had to be successfully completed in order for the proper task to begin. Instructions were automatized and included in the app. The task lasted 8-10 minutes.

Procedure

Children were tested in their usual chairs in their customary classroom, with 15 to 25 children completing the tasks simultaneously. Each child was provided with a tablet and a headset, and each was systematically checked, along with the accompanying app, before testing began. Each tablet had one of the eight lists randomly assigned to it. Each child was instructed that they would have to listen to each of the stories carefully, so as to be able to later answer questions. They were then instructed to launch the app, which began with a demo of how the response system worked along with the practice item. During the practice phase, the experimenter remained available to answer any questions or to help if necessary. Feedback on the practice item was automatically provided by the app, after which children were prompted to carry out their task at their own pace. It was made clear that no feedback or help could be provided by the experimenter during the testing phase. The metaphor task was proposed in tandem with another 5-minute-long experimental task (that is not reported here) and children were allowed to begin with the task of their choice.

Analysis

Accuracy and responses times to the metaphor task were analyzed with (generalized) mixed effect models, fitted in R (R Core Team, 2022) with the *lme4* package (Bates et al., 2015). Likelihood ratio tests were used to assess fixed effects and post-hoc contrasts were computed with the *emmeans* package (Lenth, 2022). Response time analyses were run on log-transformed data of correct answers only, after removing outliers presenting a deviation of more than 2 *SD* from each age group's mean, within each condition. Sum contrasts were used for all independent variables.

Transparency and openness

All data and analysis code are available at https://osf.io/myr2c/?view_only=7c8aeebe9837458ea5eb1c365bf20ccd.This experiment design and its analysis were not pre-registered.

Results

Before beginning our developmental analyses, we first verified that the lists, which were randomly attributed to children, were evenly distributed across age; this was statistically confirmed ($\chi^2(3) = 0.5$, $p = .66$).

Accuracy

Here, we turn to rates of accuracy on participants' picture selection. We constructed a generalized model using age group and reference condition as fixed effects, as well as their interaction, with random intercepts for items and participants. This revealed a main effect of condition (ß = 3.4, $SE = 0.25$, $p < .001$) and an interaction of condition by group (ß = 3.2, $SE = 0.45$, $p < .001$). As far as the metaphoric reference task is concerned, we observed that participants provided post-reference picture-choice accuracy rates of 43% overall along with a clear developmental pattern. Twenty-five percent of the 7-year-olds' choices in the metaphoric condition were accurate, while 63% of the older children's were (ß = -2.0, $SE = 0.28$, $p < .001$). Most errors among the metaphoric items were due to children's attraction to the pictures depicting a literal representation of the target reference (these accounted for 97% of errors, see **Figure S1.1** in supplementary materials). The two other response propositions were very rarely chosen. Contexts designed to prompt a synonymic reference mostly induced correct picture choices (at a rate of 93%). Interestingly, the contrast between age-groups was significant for this condition too but in a direction opposite of what one might expect. While practically at ceiling, the 7-year-olds committed slightly fewer errors (4%) than the 10-years-olds (10%; ß = 1.2, $SE = 0.4$, $t = 2.7$, $p < .01$). It will be seen that this result – where the 10-year-olds become wary of alternative readings in the synonymic condition – re-emerges in other findings.

To confirm that results from the task were not adversely affected by specific instantiations of the vignettes, we added *list* to the model as a fixed effect, as well as its interaction with the two other factors. This did not significantly improve the model ($\chi^2(12) = 19.6$, $p = .08$) and, importantly, it did not alter the pattern described above. Likewise, we examined results by vignette (see **Table S1.2** in supplementary materials). Item analyses confirmed that for each vignette, picture choice was determined by the preceding context. That is, accurate picture choices reflect post-metaphoric as opposed to post-synonymic reference making. More importantly, after a metaphoric reference, 10-year-olds were more likely than 7-year-olds to choose pictures accurately (as intended by the vignette's context).

For exploratory purposes, we then added *item position* as a fixed effect to the initial model, as well as its interaction with the two other fixed effects. This significantly improved the model's fit ($\chi^2(4) = 46$, $p < .001$, see **Figure S1.4** in supplementary materials), revealing no presentation order effect in the younger age group (all ps > .05), but clear effects in the older group. That is, ten-year-olds' performance improved over the course of the task with respect to the metaphoric reference condition (ß = 0.22, $SE = 0.05$, z.ratio = 4.7, $p < .001$) but they became slightly but progressively weaker in terms of accuracy in the synonymic condition (ß = -0.31, $SE = 0.08$, z.ratio = -3.8, $p < .001$). This is in line with the finding concerning 10-year-olds performance in the synonymic condition described above, where they appear to make more errors than the 7-year-olds. We will provide an account for this finding in the Discussion section.

Response times

Children's latencies while making their picture choices were analyzed with a linear mixed effect model on the correct responses. Per our pre-analytic procedure, we filtered out data that were excluded as outliers (representing 4% of the data), leaving a total of 690 data points. We constructed a linear model using the same structure as for accuracy: age group and reference condition as fixed effects, as well as their interaction, with random intercepts for items and participants.

This model revealed a main effect of age (ß = 0.19, $SE = 0.06$, $t = 3.3$, $p < .01$), as older children (mean $RT = 2966$ ms, $SD = 1534$) were faster than younger children (mean $RT = 3172$ ms, $SD = 1263$). There was also a main effect of reference condition with a slowdown induced by metaphoric contexts (ß = -0.22, $SE = 0.03$, $t = -7.0$ $p < .001$), with no interaction (ß = -0.07, $SE = 0.06$, $t = -1.1$, $p = .26$). Adding *list* as a fixed factor improved the model's fit ($\chi^2(8) = 21.1$, $p < .01$). While certain lists appeared more sensitive to effects of reference than others and while certain other lists were more sensitive to age effects, the global pattern concerning age effects and reference-condition were not unduly affected (see **Figure S1.3** in supplementary material).

Finally, we added *item position* in the task as a supplementary fixed effect to the initial model (**Figure S1.4** in supplementary materials), which also significantly improved its fit ($\chi^2(4) = 22.7$, $p < .001$). Similar to our findings on accuracy, no pattern was evident among the 7-year-olds (all $p > .05$). In contrast, the older children tended to respond faster in correctly choosing their picture as the task progressed after making metaphoric references (ß = -0.02, $SE = 0.008$, $t = -2.8$, $p < .01$) and to respond more *slowly* in correctly choosing their picture as the task progressed after making synonymic references (ß = 0.02, $SE = 0.007$, $t = 3.5$, $p < .001$).

Discussion

In Experiment 1, we tested a new tablet-based metaphor reference task on 89 typically developing children. Participants were presented with vignettes that ultimately led to the presentation of a target reference that could, depending on the prior context, be metaphoric or synonymic. Our main goal was to determine that this new task was accessible to children while producing results that resonate with prior findings.

Based on the results of the current Experiment, we can draw the following five conclusions. First, given children's performance with the control items (the synonymic-reference condition), one can be confident that young participants are carrying out the task competently. Error rates were low overall for both age groups. Second, the data show that children were sensitive to preceding context, providing evidence for the task's face validity. Vignettes that prompted the listener to make a metaphoric reference led to more equivocality among all the children. That is, the vignettes that were designed to prompt metaphoric references were also prone to errors (overall, depictions of a literal representation of a metaphor target reference were chosen roughly half the time). Third, and most importantly to our study, rates of accurate (intended) metaphoric choices increase with age: the 10-year-olds were more likely than the 7-year-olds to choose pictures that show that they made metaphoric references. These observations also applied to vignettes taken individually. Fourth, the post-metaphoric-reference picture choices come with a cost in terms of response times when compared to post-synonymic-reference picture choices that follow from the same target expressions. This makes sense for the simple reason that, generally speaking, metaphoric picture-choices were juxtaposed with tempting literal depictions of the target reference. Fifth, the addition of finer exploratory analyses – such as those that add list effects or trial effects – only enhance the main results. Overall, these results confirm the task's viability as well as its ability to lay bare a developmental sensitivity of metaphor comprehension over the school age years.

The 10-year-olds in particular produced interesting outcomes. Among other things, we observed that performance changed over the course of the task for this age group. Notably, the 10-year-olds' picture choices made in the wake of the metaphoric-reference target sentence were carried out with both greater accuracy and decreased latency over the course of the experimental session. Interestingly, we also observed that there was a slight negative effect with respect to their picture choices in the wake of the synonymic target-reference. The 10-year-olds were slightly, but significantly, more likely to consider alternatives even to the synonymic reference; likewise, the 10-year-olds took longer to correctly choose the picture that corresponds to the literal representation of the target reference, as the task wore on. We surmise that, based on their experience with the task and its multiple metaphoric items, children progressively became aware of dual meanings in the task and that this prompted them to be more discerning, even among the vignettes that were designed to have a univocal interpretation. Indeed, errors in the synonymic condition that augmented in the course of the task reflected metaphorical interpretations of the reference. Generally speaking, this ability to appreciate an alternative meaning captures well how 10-year-olds differ from their younger cohorts.

EXPERIMENT 2

Experiment 1's results allow us to conclude that the metaphor referential task here provides the kind of outcomes one would expect based on prior studies. That is, metaphor-reference abilities appear to become more adult-like with age. Likewise, one generally finds metaphor-related slowdowns, which also resonate with the previous literature. In the present case, the results are based on the way participants make their choice of pictures (and not, for example, on reading times).

In Experiment 2, the same task is used to further analyze the developmental effect with the additional aim of determining the extent to which two different growing abilities – linguistic competence and TOM – contribute to children's metaphoric performance. That is, we investigate interindividual differences in children's metaphor comprehension in terms of age, like in Experiment 1, and this time while also considering other cognitive abilities, which can be used as predictors in analyses. In order to provide this sort of assessment, the task is presented in the context of a standardized linguistic-pragmatic assessment battery. This is why it is presented in a single order and while interdigitated with other pragmatic tasks. In other words, the metaphor reference task here is similar to the one in the prior experiment, except that it is embedded in a larger and longer exercise, which includes in effect filler items. Assuming that we replicate the findings from Experiment 1, it would only increase the metaphor reference task's validity.

***Table 1**: Experiment 2's sample characteristics by age groups*

	6 YO	7 YO	8 YO	9 YO	10 YO	Whole sample	Age group difference
N	53	59	47	41	48	248	/
Proportion female	43%	54%	45%	56%	58%	51 %	$\chi^2(4) = 3.7$, $p = .45$
Proportion from school A	54%	54%	36%	58%	50%	51%	$\chi^2(4) = 4.6$, $p = .33$
Mean SES Index (SD)[1]	59 (15)	59 (18)	57 (16)	62 (17)	56 (16)	58 (16)	$F(4,244) = .85$, $p=.50$

[1] Data were missing for 3 participants who were removed from these analyses.

Method

Participants

Two hundred and forty-eight children took part in this study. Children were between 6;0 to 10;11 years old (mean = 8;4 years old, *SD* = 1;5) and were homogeneously distributed across age groups (see **Table 1**). As in Experiment 1, all participants regularly attended an elementary school (*école primaire*) and were native speakers of French. Both parents and children provided informed consent. As declared by parents, children had no neurodevelopmental disorder (learning, autism spectrum, attention-deficit, or language-related disorder), nor any sensory (visual or auditory) or motor disorders that would prevent them from using a tablet.

Language in general is much associated with childen's socio-economic environment (see, e.g., Di Sante & Potvin, 2022). To provide social balance, children were thus recruited from two different French schools, one private school in an urban economically privileged area (School A) and one public school situated in a rural and less economically privileged region (School B). Moreover, we collected, via parental questionnaires, an estimation of family income through the Family Affluence Scale (FAS, Currie et al., 2008), which is based on relatively non-intrusive questions and has proven to be a good indicator of family wealth across countries (Boyce et al., 2006; Currie et al., 2008). We also asked parents to provide their education levels, which were coded on a scale ranging from 0 (no diploma) to 7 (PhD). For each child, parents' mean education level is used when information about both parents is available; when data from only one parent is available (6 % of children), the score of that one parent is reported. The FAS and the parental education level are standardized then averaged to provide a socio-economic status composite (SES) varying from 0 to 100.

This study was authorized by the two schools where children were included and tested. It was part of a project which also required the inclusion of participants with neurodevelopmental disorders (not reported here), and as such received ethics approval from the [name of the board and ID RCB*, masked for review*].

Materials

Metaphor task. This Experiment employed a unique version of the task by selecting one of the lists from Experiment 1. The advantage of preparing the task in this way is that potential differences between two children cannot be attributed to list or presentation order effects. Moreover, this approach will allow us to consider the data collected as normative for individual subjects, which will subsequently permit us to make clinical applications while also extending our research on interindividual differences to atypical development.

As part of the unique version employed here, we mixed the 12 items from the first list of Experiment 1 with 22 filler items of similar structure and response format which involved other language and cognitive abilities such as memory, irony or conditional reasoning. These are not of interest to the present paper and will not be discussed further. For this version, the task is preceded by 3 training items to familiarize children with the response format, which relied on literal material only and provided the participant with automatic feedback. To sustain engagement and attention for this longer version of the task, two breaks were included at approximately a third of the way through and then at three-quarters of the way through, at which points participants viewed 30 seconds of non-verbal video-clips of funny animals. The task lasted approximately 20 minutes.

As was the case for Experiment 1, we verified that in the selected list, the vignettes that set up the metaphoric and synonymic references remained matched for (1) length, (2) readability, (3) mean frequency of stories content words (nouns, verbs, adjectives, adverbs) in an oral French corpus, and frequency of both (4) referent and (5) target nouns in school textbooks (see **Table S2.1** in supplementary materials).

Parental appreciation of pragmatic abilities. To provide external validation for the metaphor task, we used a subset of the validated parental questionnaire of children's communicative abilities developed by Bishop (2003), the French adaptation of the Children's Communication Checklist-2 (Vézina et al., 2013). In this questionnaire, parents are asked to judge the frequency at which they observe different communication-related behaviors in their child. We extracted the items from the "Use of context" subscale which targets context-dependent uses of language such as figurative language, including metaphors (e.g., "How often is your child over-literal, sometimes with (unintentionally) humorous results?"). As a control, we used the items from the "Semantics" subscale, which targets non-pragmatic linguistic behaviors, such as word finding (e.g., "How often does your child make false starts, and appears to grope for the right words?"). Each subscale is composed of 7 items and provides a total score ranging from 0 to 21, with higher values being associated with higher communication difficulties.

Grammar task. Grammar reception was assessed with the *BILO-3C Oral comprehension* task (Khomsi et al., 2007). This is a computerized French-validated picture-matching task targeting sentences with different morpho-syntactic phenomena such as verbal inflexions of number and tense, passive structures, object pronouns, direct and indirect relative clauses. For each item, four pictures are displayed on a screen, and children are asked to select the picture that best-matches an aurally presented sentence. For example, to assess passive structures, children are aurally presented a sentence, such as *Marie est poussée par Pierre* (*Marie is pushed by Pierre*), while being shown four pictures: i) a boy pushing a girl on a bicycle (expected answer), ii) a girl pushing a boy on a bicycle, iii) a boy watching a girl on a bicycle, iv) a boy and a girl pushing a stroller). It includes 2 training items and 27 test items, and provides a total score ranging from 0 to 27.

Theory of mind task. TOM was assessed with a tablet-based version of the Picture Sequencing Task (PST), an experimental task designed by Langdon & Coltheart (1999). In this minimally-verbal task, participants are asked to sequence four pictures. Following guidelines developed by Langdon & Coltheart (1999), participants receive 6 points for properly positioning all of the pictures of each sequence (2 points for each of the first and last picture, and 1 point for each of the intermediate pictures). Following Rajkumar et al. (2008), we used a subset of 12 items from Landgdon & Coltheart's (1999) original material, broken down evenly across three conditions. That is, four sequences involve mechanical causalities (e.g., consider pictures showing how a speeding truck's vibrations prompt a boulder to roll down a hill), four were based on social scripts (e.g., consider a sequence that represents meeting a friend for a coffee) and four critical sequences involve false beliefs (as described above, these involve detecting a false attribution). There were also two training items. Scores for the first two control categories are averaged to provide a General Sequencing Abilities (GSA) index, which has proven to be strongly correlated with IQ among typically developing school-aged children (Rajkumar et al., 2008). This measure is used as a control measure while the average of the false-belief scores constitutes the TOM index. Both indexes vary from 0 to 6. The PST has been used to study TOM skills in various populations, including high vs. low schizotypal healthy participants in the original study (Langdon & Coltheart, 1999), children and adults with William's syndrome (Porter et al., 2008) as well as adults with bipolar disorder (Van Rheenen & Rossell, 2013). The tablet-PST's overall Cronbach's alpha was .76 in our sample (for more information, see Petit et al., *under review*).

Procedure

Parental consent forms and questionnaires (inclusion criteria check, SES, CCC-2) were collected before testing. As in Experiment 1, assessments took place in the children's usual classrooms. Children were first tested with a portion containing the metaphor task; this took roughly 20 to 30 minutes. TOM and language assessments took place in a second session on a different day. This second day of testing lasted roughly 20 to 30 minutes as well.

Analysis

Before analyzing the data, we took two preliminary steps. First, as this study is targeting typical development only, we removed those participants whose rates of correct responding were 2.5 SD below their age group's mean in the grammar or TOM task. Second, for response time analyses, a) the data were log-transformed and b) incorrect responses and outliers were removed based on the procedure used in Experiment 1. Participants with missing data for SES were also excluded from the analysis using this variable. All analyses were performed in *R* (R Core Team, 2022).

We begin by testing the structural and external validity of the metaphor-reference task. This is first done by assessing its 2-factor structure (metaphoric vs synonymic reference) with a confirmatory factor analysis (CFA), using the *lavaan* package (Rosseel, 2012). Then, we examined Cronbach's alpha for each of these 2 conditions. To determine external validity, we computed correlations between accuracy in the critical condition of the metaphor reference task with i) the CCC-2 "Use of context" subscale, to provide convergent validity (where a correlation is expected), and ii) the CCC-2 "Semantics" subscale to provide divergent validity (where no correlation is expected).

Then, we fitted mixed effects models, i.e., generalized mixed effect models (for accuracy) and linear mixed effects models (for response times). These were fitted with the *lme4* package (Bates et al., 2015). In order to address our research concerns, we followed the following two steps. First, we fitted hierarchical models that include reference condition as well as single, quadratic and then cubic terms for age as fixed effects, as well as their interactions with condition. Random structure included random intercepts for participants and items. This should allow us to describe the precise evolution of performances with age. Second, with the aim of assessing the role of the different cognitive predictors on the DV, we dropped the higher order terms for age, to prevent computation as well as interpretation issues. Then we added the predictors of interest as fixed effects (grammar and TOM, as well as the GSA, to control for sequencing abilities) and allowed for 3-way interactions of these predictors with age and condition. In that way we could describe each predictor's effect across reference conditions and at different ages. SES was also added as a control covariate. The final model formula was thus *DV ~ SES + age * condition * (Grammar + GSA + TOM) + (1|participant) + (1|item).*

In each model, age was entered in years as a continuous predictor (this follows Royston et al. [2006] who argue that continuous predictors should not be dichotomized). For descriptive purposes, and because interactions involving multiple continuous predictors are complicated to examine, effects at different ages were described via post-hoc contrasts run from the fitted models at the values 6, 7, 8, 9 and 10 years of age (as could have been done for any particular value) and through visualizations. This should not be taken to imply that we used age-groups as a categorical variable in the analysis. Sequential difference coding was used as a contrast for age and sum contrast for reference condition. All continuous predictors were centered on their mean with the *scale()* function. Models' assumptions were checked before reporting their estimates. Single effects

Table 2: *Final sample (N=236) results on the control tasks by age groups, and correlations with age*

	Groups	6 YO N = 49	7 YO N = 55	8 YO N = 45	9 YO N = 41	10 YO N = 46	*Correlation with age*
Grammar	Mean	13.2	14.5	16.4	17.9	18.3	*r = .51*
	(*SD*)	(3.53)	(2.98)	(3.71)	(2.92)	(3.50)	*p < .001*
TOM	Mean	2.90	3.43	4.13	4.54	4.02	*r = .37*
	(*SD*)	(1.01)	(1.17)	(1.24)	(1.10)	(1.26)	*p < .001*
GSA	Mean	4.44	4.77	4.94	5.47	5.27	*r = .36*
	(*SD*)	(1.19)	(0.82)	(0.74)	(0.56)	(0.80)	*p < .001*
CCC-2 Semantics	Mean	6.6	6.0	6.3	6.2	6.1	*r = -.0*
	(*SD*)	1.7	1.6	1.7	1.7	1.4	*p = .23*
CCC-2 Context	Mean	5.3	4.9	5.1	3.5	3.9	*r = -.21*
	(*SD*)	2.7	2.6	3.4	2.5	2.5	*p < .01*

were assessed with likelihood ratio tests. All post-hoc contrasts were performed with the *emmeans* package (Lenth, 2022) while adjusting for multiple comparisons.

Transparency and openness

All data and analysis code are available at https://osf.io/myr2c/?view_only=7c8aeebe9837458ea5eb1c365bf20ccd.This experiment design and its analysis were not pre-registered.

Results

In the sample, 10 participants out of 248 (4%) scored below 2.5 *SD* from their age-group mean on the control tasks and were thus excluded. One participant did not complete the control task and was also excluded, leaving a remaining sample of 237 children for the analysis. For descriptive purposes, **Table 2** presents the distribution of children across age-groups and their results on control tasks.

Task validation

The confirmatory factor analysis clearly confirmed the bifactorial structure of the metaphor task, one that contrasts synonymic and metaphoric items (*CFI* = .98, *TLI* = .98, *RMSEA* = .03), with latent factors sharing no variance (ß = 0.001, *SE* = 0.001, *p* = .49). The internal consistency of the metaphorical condition proved to be excellent without revealing redundancy in the items (α = .81, 95% *CI* = [.77-.84]); meanwhile, the synonymic condition appeared to provide less consistency (α = .50, 95% *CI* = [.41-.60]), which could be accounted for by the ceiling effect observed in this condition (see below). Due to this ceiling effect, convergent and divergent validity was assessed only for the metaphor condition. For the 225 children whose parents filled the entire questionnaire, the individual mean accuracy rate for metaphoric reference was significantly correlated with the parental report of pragmatic abilities (CCC-2 "Use of context" r = -.21, p < .01) but not with semantic abilities (CCC-2 "Semantics", r = .02, p = .81). Finally, the results of experiment 2 were very consistent to those of experiment 1: the seven- and 10-year-olds accurately choose the metaphoric representations at rates of 29% and 61%, respectively in this Experiment, contra 25% and 63% in Experiment 1.

Accuracy

Accuracy in the synonymic condition was very high overall (95%), indicating that participants correctly chose the intended target (a picture reflecting a literal representation of the reference) here. As expected, and as shown in **Figure 2**, rates of correct picture choices in the metaphoric-reference condition were lower

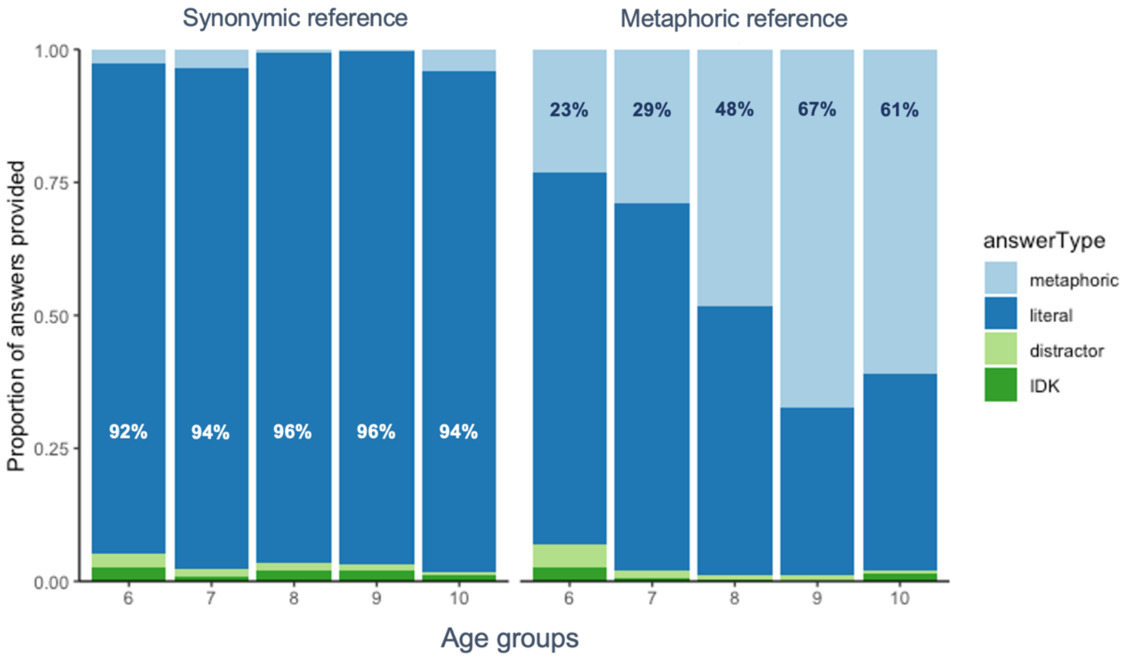


Figure 2: *Proportion rates of each answer for the different conditions (based on raw data), across the different age groups, with accuracy for each age group in each condition*

Table 3: Outcomes of the generalized mixed effects model explaining accuracy in the task (left part, orange) and of the linear mixed effects model explaining log-transformed response times to the picture selection portion of the task (right part, green)

	ACCURACY			LOG(RT)		
PREDICTORS	OR	95% CI	p	Estimates	95% *CI*	P
(INTERCEPT)	5.10	3.49 – 7.47	<0.001	8.25	8.03 – 8.47	<0.001
SES	1.20	0.98 – 1.47	0.072	**-0.04**	**-0.07 – -0.00**	**0.023**
AGE	**1.50**	**1.16 – 1.94**	**0.002**	**-0.04**	**-0.06 – -0.01**	**0.008**
CONDITION	**0.01**	**0.01 – 0.03**	**<0.001**	**0.34**	**0.05 – 0.62**	**0.022**
GRAM	1.16	0.88 – 1.52	0.304	-0.06	-0.27 – 0.14	0.542
GSA	1.12	0.86 – 1.46	0.409	-0.08	-0.26 – 0.10	0.387
TOM	1.00	0.78 – 1.28	0.989	**-0.22**	**-0.44 – -0.01**	**0.043**
AGE * CONDITION	**1.86**	**1.28 – 2.71**	**0.001**	-0.02	-0.05 – 0.01	0.231
AGE * GRAM	1.11	0.87 – 1.42	0.380	0.01	-0.02 – 0.03	0.472
AGE * GSA	0.91	0.72 – 1.15	0.441	0.01	-0.01 – 0.03	0.428
AGE * TOM	0.99	0.78 – 1.27	0.966	0.02	-0.00 – 0.05	0.067
CONDITION * GRAM	1.69	1.12 – 2.54	0.012	0.24	-0.02 – 0.50	0.074
CONDITION * GSA	**0.66**	**0.45 – 0.97**	**0.033**	-0.20	-0.44 – 0.04	0.103
CONDITION * TOM	**1.51**	**1.04 – 2.20**	**0.029**	**-0.36**	**-0.64 – -0.07**	**0.013**
(AGE * CONDITION) * GRAM	**2.31**	**1.62 – 3.28**	**<0.001**	**-0.03**	**-0.06 – -0.00**	**0.034**
(AGE * CONDITION) * GSA	1.14	0.81 – 1.60	0.453	0.02	-0.01 – 0.05	0.168
(AGE * CONDITION) * TOM	**0.71**	**0.50 – 1.00**	**0.05**	**0.04**	**0.01 – 0.07**	**0.018**
RANDOM EFFECTS						
Σ^2	3.29			0.12		
T_{00} PARTICIPANT	1.19			0.03		
T_{00} ITEM	0.26			0.01		
ICC	0.31			0.25		

(44%). Also, as expected, most errors resulted from the selection of the pictorial representation of the reference taken literally (this accounted for 95% of errors).

The first-step analysis revealed that the cubic model outperformed the linear model ($\chi^2(4)$ = 16.9, $p < .01$). This model confirmed that participants were much more likely to succeed the synonymic-reference items than they were in the metaphoric-reference items (*OR* = 4.1, *SE* = 0.38, $p < .0001$). Post-hoc contrasts revealed no age effect in the synonymic condition. As for the metaphoric-reference condition, the 6- and 7-years-olds prompted the lowest accuracy rates, the 8-year-olds scored higher than the youngest participants (ß = 1.2, *SE* = 0.2, $p < .0001$), and the 9-year-olds outperformed the 8-year-olds (ß = 1.2, *SE* = 0.2, $p < .01$). The 10-year-olds' rates of accuracy were ultimately comparable to those of the 9-year-olds ($p > .05$).

In a second step, we dropped the cubic and quadratic terms for age and added the additional predictors (see model's output in **Table 3**). This model revealed that accuracy was overall more predicted by both TOM and grammar in the metaphoric than in the synonymic, reference condition. These variables and the reference condition were involved in 3-way interactions with age: TOM appears to have a positive effect on picture choice accuracy in the metaphoric reference, as opposed to the synonymic reference condition, but the influence of TOM decreased with age. As for the role of grammar, one sees the reverse pattern, i.e., it reveals having an *increasing* influence on metaphor accuracy with age, compared to the synonymic condition. Of the two, grammar appears to be more impactful on the three-way interaction than TOM is.

The predicted marginal effects of TOM and language on accuracy in the reference task, as a function of age, is plotted in **Figure 3**. Post-hoc contrasts (see **Tables S2.2** and **S2.3** in supplementary materials) revealed that higher scores on TOM predicted a higher accuracy on the metaphoric reference task compared to the synonymic controls at 6, 7 and 8 years of age, but not at 9 or 10 (see Figure 3, upper half). Conversely, higher grammar abilities predicted higher rates of accuracy in the metaphoric condition compared to the synonymic condition at 8, 9 and 10 years old, though not for the 6 and 7-year-olds (Figure 3, lower half). Interestingly, the only children in the entire sample to reach accuracy rates for the metaphoric-reference items that were comparable to those of the synonymic-reference items were those 10-year-

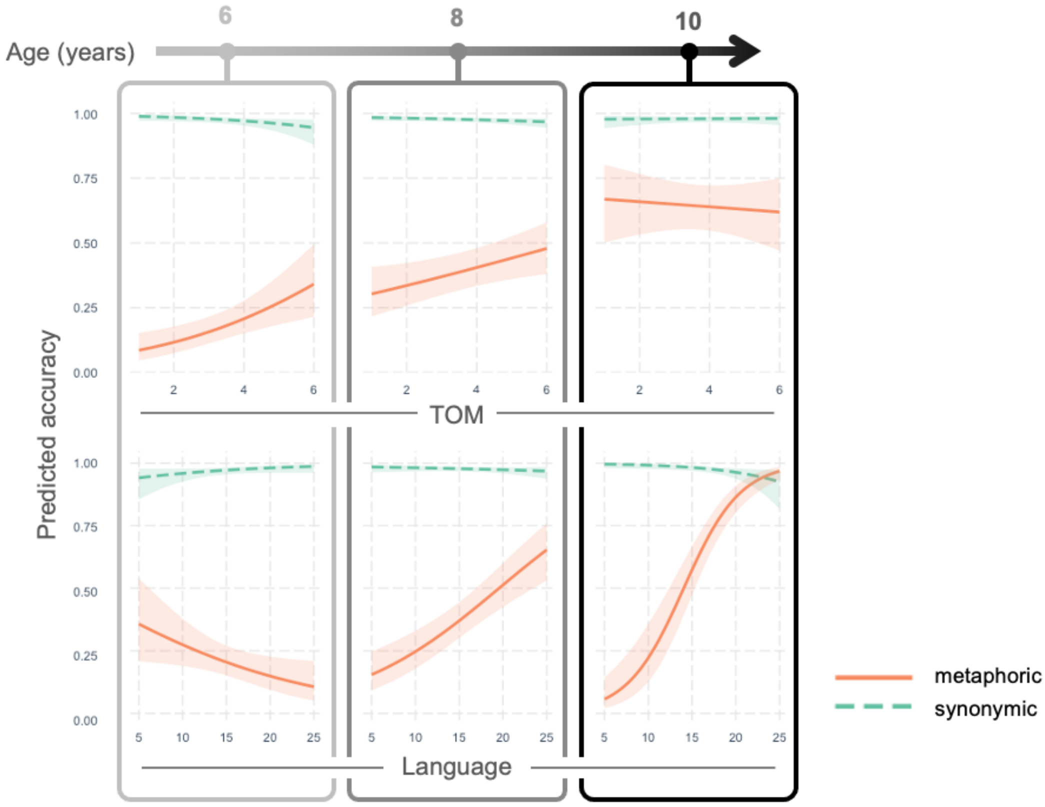


Figure 3: *Predicted accuracy in the different conditions, as a function of TOM (x-axis, above) and grammar (x-axis, below), at 3 different ages (6-, 8- and 10-year-olds) (predictions were backtransformed from scaled variables)*

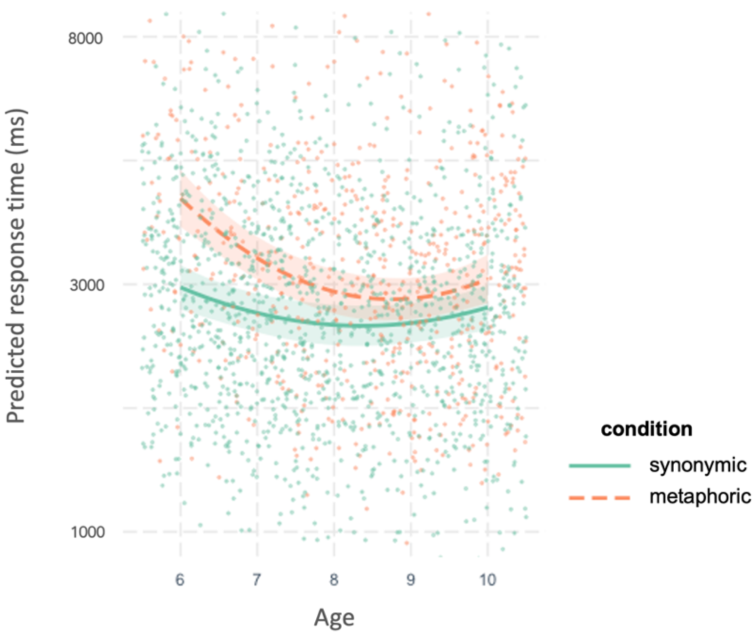


Figure 4: *Predicted response times (in ms, on a logarithmic scale) as a function of age and condition, and partial residuals (residuals below 1000 ms or above 8000 ms [i.e., 1% of the data points] were not displayed and age was backtransformed from scaled values*

olds who scored highly on grammar (see Figure 3, bottom right).

We note that the TOM index and language total score were themselves clearly correlated ($r = .49$, $p < .001$, see correlation matrix of all predictors in **Figure S2.1** in supplementary materials), but that the risk of multicollinearity for the model was limited (all *VIF* values < 2.33).

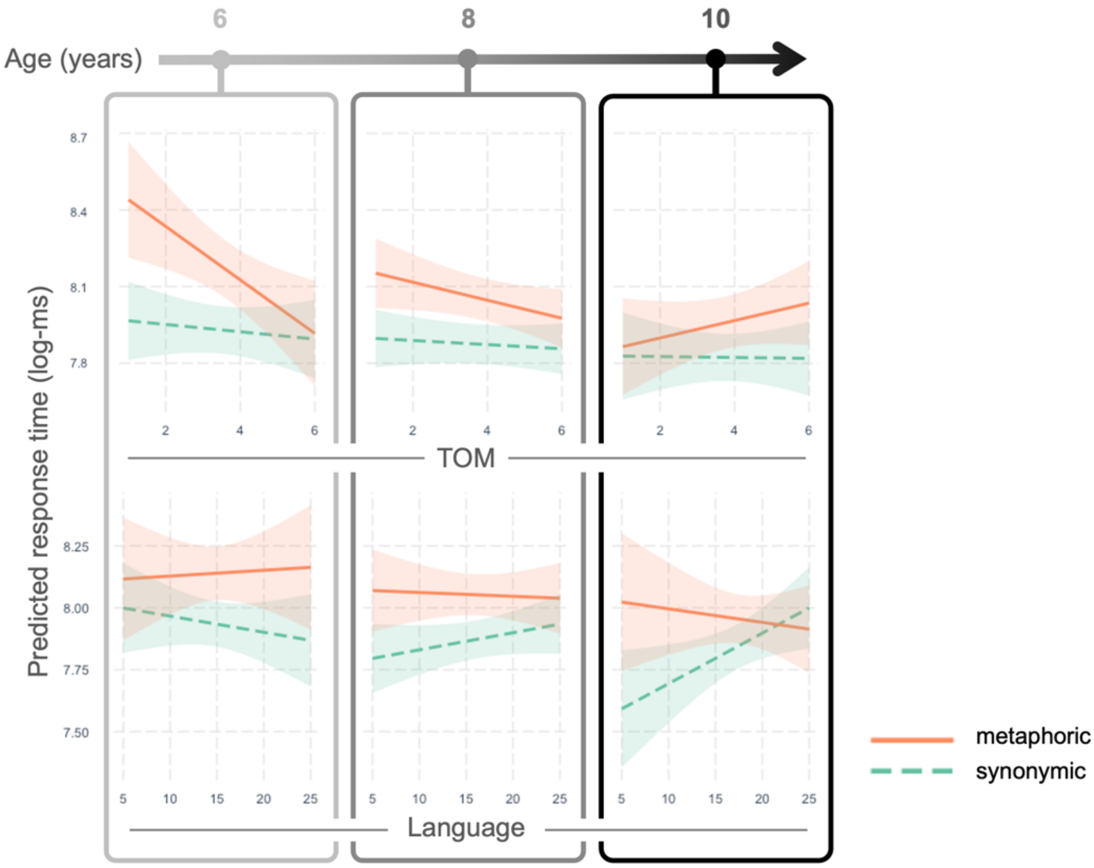


***Figure 5**: Predicted response times (in ms, log-transformed) across the two reference conditions, as a function of TOM (x-axis, top half) and grammar (x-axis, bottom half), at 3 different age points (6, 8 and 10 years old) (predictions were backtransformed from scaled values)*

Response times

Before running the RT analysis, we filtered the data set so that it included correct responses only (this amounts to removing 31% of the data points at this stage). We then removed outliers (4.6% of the remaining responses). The final set included 1880 log-transformed data points. In the first step, the quadratic model proved to provide a better fit to the data than the linear one ($\chi^2(2) = 22.5$, $p < .001$), but adding a cubic term did not improve the fit ($\chi^2(2) = 0.8$, $p = .70$). As can be seen in **Figure 4**, the selected quadratic model revealed a main effect of condition, with accurate post-metaphoric picture choices prompting longer response times compared to post-synonymic ones (ß = 0.18, *SE* = 0.05, $p < .01$). The estimated mean picture choice time for the post-synonymic control items was 2527 ms (*SE* = 105) as opposed to 2879 ms (*SE* = 131) for the metaphoric ones. We will refer to the difference between these two conditions as the *metaphor-related cost.*

Reference condition was part of an interaction with age, so one can see that the metaphor-related cost evolved with age. Post-hoc tests revealed that the metaphor-related cost decreased between 6 and 7 years of age (ß = 0.14, *SE* = 0.04, $p < .001$) as well as between 7 and 8 (ß = 0.09, *SE* = 0.02, $p < .0001$) and marginally between 8 and 9 years of age (ß = 0.04, *SE* = 0.02, $p = .07$) but it did not change between 9 and 10 ($p = .88$).

In a second step, we dropped the quadratic term for age and added the other covariates. The model's output is reported in **Table 3**. The model's estimates indicate that response times are involved in three-way interactions of TOM and grammar with age and condition. Higher TOM abilities are associated with faster responses after making a metaphoric reference as opposed to a synonymic one, but this influence decreased with age. The opposite pattern was observed for grammar, as for accuracy.

Post-hoc contrasts (see **table S2.4** in supplementary materials) revealed that higher TOM predicted faster pictorial choice responses after making metaphoric rather than synonymic references at 6 and 7 years of age, but that no such effects were significant at older ages

(**Figure 5**, upper half). Grammar was linked to response times of metaphoric as opposed to synonymic reference pictorial choices at 9 and 10, but not before. Visual inspection of the predictions (**Figure 5**, lower half) as well as estimates (**table S2.5** in supplementary materials) indicate that this effect is driven by a negative impact of language score on post-synonymic picture-choice response times at these ages (leading to longer response times), rather than being due to a positive impact of language on post-metaphoric picture-choice response times.

As far as how underlying factors affect the development of metaphor comprehension is concerned, **Figure 6** summarizes the marginal effects of TOM and language as a function of age on accuracy (left) and response times (right) in the metaphoric reference condition. One can see that TOM has a weakening influence on both accuracy and Picture-choice response times with age. Meanwhile, language competency has a facilitating effect on accurate metaphor picture choice with age.

Exploratory analysis

To further explore the data and to characterize the nature of the post-reference picture choices, we fitted a supplementary model on the developmental trajectory model of response-times by adding – as a third response-type – responses that reflect the literal picture choice in the metaphor reference condition, which were previously considered as errors. In other words, we compare three kinds of response times: a) those that concern correctly choosing pictures among post-synonymic items, b) those that concern correctly choosing pictures in the wake of a metaphoric reference and; c) those that concern incorrectly choosing literal representations of the target reference pictures in the wake of the metaphorical-reference. This model showed that so-called *literal* readings of a metaphoric target reference were made more quickly than those that reflect correctly choosing pictures post-metaphorically, regardless of age (all ps < .001) and. Moreover, these incorrect choices were made as fast as those found among correct post-synonymic items (all ps > .05, see **Table S2.6** and **Figure S2.3** in supplementary materials). This indicates that the choices reflecting literal interpretations in the metaphor reference condition were made precipitously and not ponderously.

Discussion

In Experiment 2, we tested the tablet-based metaphor reference task on a larger group of children while presenting it as part of an assessment-like scenario. As anticipated, the results with respect to metaphor development here were consistent with those found in Experiment 1. Moreover, it provided finer detail since it included participants across the entire age range, i.e., between 6 and 10 years of age. We were thus able to show that metaphor comprehension development is non-linear and especially focused between 7 and 9 years old.

More germane to our investigation was determining the extent to which the

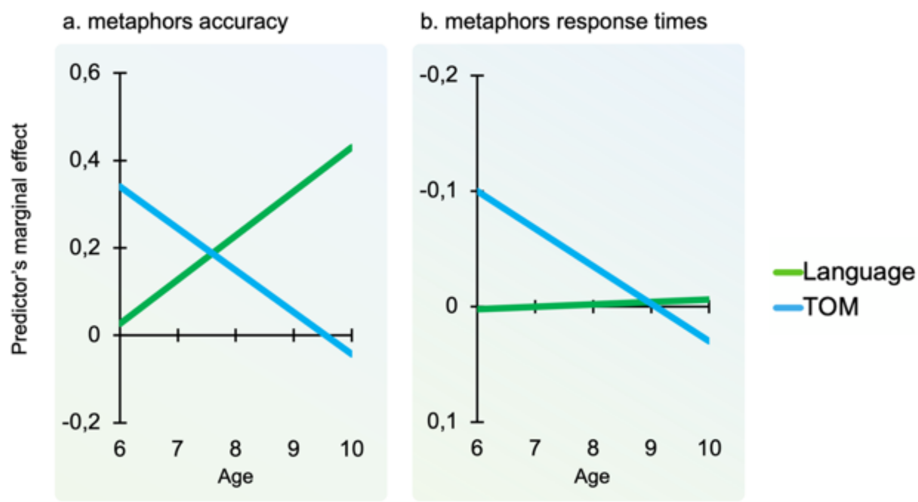


***Figure 6**: Predicted partial effects of TOM and language on accuracy (left) and response times for correct responses (right) in the metaphoric reference condition, as a function of age (note that the y axis in (b) reverses polarity to indicate increased speed).*

development of TOM and grammar each appears to interact with metaphor development. Interestingly, the two cognitive abilities indeed appear to interact with development and in two different times in the chronological age of the participants. Theory of mind appears to have facilitative effects on making accurate metaphoric choices among the younger children. However, the effects of theory of mind on metaphoric choices plateau by the time children are 9 years old. In contrast, grammar competence appears to have facilitative effects on the older children. Indeed, the only children to choose pictures correctly after a metaphoric reference at rates as high as the synonymic reference were the oldest children who scored highest for grammar.

The role of the two cognitive abilities with respect to reaction times was similarly revealing. For those children who accurately make metaphoric picture choices, we reported that the metaphor-related cost evolved with age. That is, the difference in metaphor-based and literal-based picture choice responding decreased between 6 and 9 years of age but it did not change between 9 and 10. Grammar was associated to response-time differences between metaphoric and synonymic references after 9 years of age. Much like in Experiment 1, this appears to be due to the fact that increased sophistication with respect to language impacts post-*synonymic* picture-choice response times in this age range (leading to longer response times in the control condition). That is, the older participants appear to slow down in the synonymic condition as they become aware of alternative meanings for the target reference rather than becoming faster due to a positive impact of language on post-metaphoric picture-choice response times.

Overall, one can see almost a hand-off between the two cognitive substrates. Early on, TOM is influential on picture choice in the metaphoric reference condition among the younger participants. This appears to wane by the time children are roughly 8 or 9 years old. At this point, grammar competence appears to take on an increasing influential role on children's metaphorical responding.

General Discussion

We began this paper by providing background on prior metaphor development studies and by introducing a tablet-based referential metaphor task. In line with previous studies, data from Experiment 1 provided evidence indicating that metaphor comprehension develops with age and showed that metaphor-based picture selection, which reflects the study's dependent variable, took reliably longer than its controls. Importantly for the present paper, the task in Experiment 1 provided us with a proof of concept of the metaphor task's viability before it was juxtaposed with grammar and TOM tasks in a second experiment, that was much more comprehensive. That is, in Experiment 2, we included the metaphor-reference task as part of an assessment of pragmatic abilities that was distributed to a large sample of 238 typically-developing children, along with a non-verbal TOM task and a receptive grammar task. This second study allowed us to characterize the development of, and arguably the cognitive predictors of, referential metaphors. Not surprisingly, the data from the metaphor task in Experiment 2 was highly similar to those in Experiment 1. In the remainder of the General Discussion, we consider in greater detail what the task and the Experiments here bring to the rich literature on metaphor development. This detailed examination addresses the two general questions that motivated this research, one concerning the added value of the new variation of the task and another concerning what the adjacent cognitive tasks – on TOM and receptive grammar – reveal about their influence on metaphor comprehension development.

What does the current metaphor reference task provide for the literature?

The developmental trend that appeared in both Experiments is consistent with a series of results indicating that school-age is a key period for metaphor development. That is, all the children had little difficulty providing correct responses in the synonymic-reference control condition (which requires participants to choose the single picture that corresponds with the literal meaning of the target reference). This highlights how the linguistic material and the response format of our task is accessible to children. Meanwhile, picture choice-making in the metaphorical reference condition revealed greater equivocality. The data showed that incorrect picture choices in the metaphoric-reference condition typically reflects *literal* interpretations of the target reference. As children become older, they show improved metaphor comprehension performance, especially between the ages of 7 and 9 years, and without particular floor or ceiling effects. Up to about 8 to 9 years of age, the effort needed to choose metaphoric-based pictures as opposed to literally-based ones, in terms of response times, largely diminishes before stabilizing.

Although the developmental trend described in our Experiments is similar to those found in previous studies (Noveck et al., 2001; Seigneuric et al., 2016; Tonini et al., 2023), our rates of accuracy were somewhat lower. This can be explained by three methodological modifications made to the present study that are not found in the prior ones. First, we tested younger children while using oral, instead of written, materials. This has the advantage of avoiding interference with literacy because our youngest participants most likely are not accomplished readers. Second, we used a multiple-choice response format (instead of, say, questions requiring yes-no responses), which yields more variability. Importantly, this also allowed us to extend previous results by providing literal interpretations of metaphoric references as a picture choice option (see Van Herwegen et al., 2013) and to positively identify a large proportion of participants' errors as non-random and literal. Third, we designed and selected metaphoric references that could then be made part of sensible initial contexts. Likewise, we built sensible contexts around our synonymic target references. In other words, by building up contexts around a sensible metaphoric or synonymic reference (rather than keeping a single context in which two different references are used), we were able to come up with highly reliable vignettes.

This study is also different from prior studies with respect to the way it records latency measures, when they exist. In prior studies, the dependent variable was the time needed to read a critical sentence that could contain a metaphoric or a literal reference. These perhaps directly grasp the effort associated with the online processing of the reference, but they could arguably be confounded with other reading-related procedures. In our task, we collect response times that occur downstream from the target reference. Interestingly, we still report metaphor related slowdowns. This shows that a) effects related to metaphor slowdowns are robust and that; b) it might not be related to referring back to a previously mentioned element. It could be the case that, for successful metaphor comprehenders, it is not reference-making in itself that is difficult but rather how to proceed once the metaphoric reference arrives. The current study appears to put the onus on what happens once a metaphorical reference arises. Future studies could further examine where the source of difficulty in a metaphoric reference task lies. Note that we would not necessarily hypothesize that metaphor-related slowdowns are merely costly; they are also argued to come with benefits by helping encoding and memorization (Noveck et al., 2001; Reynolds & Schwartz, 1983).

From the point of view of the current study's qualitative results, one finds consistency with both classical (Winner et al., 1976) and more recent studies (Deckert et al., 2019; Lecce et al., 2019). The current findings on age-related advances are also important because they are coherent with prior observations from other types of metaphor tasks. Our data are also in line with Willinger et al. (2019) and Deckert et al. (2019), who suggest that there are spurts in metaphor development among school aged children, as opposed to continuous development.

To what extent do theory of mind and language abilities provide scaffolding for metaphoric reference comprehension?

We now turn to our second research question, which concerns identifying the potential role played by TOM and grammar in metaphor development. Before we begin, recall two features that served as the basis of our analyses. One is that we adopted a minimally-verbal measure of TOM (from Langdon & Coltheart, 1999) for the purposes of the current study. This is unlike TOM measures in other studies in this literature, which typically employ TOM assessments that are verbally rich. Consider the oft-used Strange Stories task (Happé, 1994), used in Lecce et al. (2019), which asks participants to detect subtle differences between jokes, white lies, misunderstandings and so on and through rather long vignettes and questions.[1] These sorts of tasks make it potentially difficult to disentangle the respective roles of language and theory of mind, which is central to our goal here (see de Villiers, 2007 on the interface between language and TOM). The PST-based measure advances our understanding of the three-way relationship between metaphor comprehension, language and TOM abilities, as called for by Matthews et al. (2018). The other is that we used age as a continuous predictor in our statistical analysis, in interaction with the other variables, which allows us to exploit the whole sample's size and variability. This implies that the turning points we report should be considered as qualitative indications, rather than as exact values.

Our results reveal that TOM was positively associated with both accuracy and response times of picture choices after having heard metaphorical references as opposed to synonymic one. This TOM effect is clear at 6 years of age and steadily decreases with age, falling below significance threshold at roughly 9 years of age. This finding is in line with the hypothesis proposed in Lecce et al. (2019)'s study, which relied on tasks in which children metalinguistically explain a metaphor's meaning, and with Tonini et al. (2023) who, in addition, used a referential task. The current study shows that this observation can arguably be generalized because our data are generated by referential metaphors, tested with a multiple-choice format, and when TOM is assessed with our minimally verbal task. Hence, one can confidently conclude that the association we and others report between TOM and metaphor development is not linked to a language-based confound. Our paradigm also allowed us to test younger children and without reading, providing evidence that TOM's influence can be observed in children who are as young as 6-years-old. That said, the work here confirms and extends the hypothesis that says TOM has a passing influence on metaphor comprehension.

Formal language skills, in the form of receptive grammar, also revealed themselves to be statistically predictive of metaphorical reference behavior, even though this association was visible with respect to accuracy only (i.e., not to picture choice response times). This generally points to an important developmental shift during school age for the comprehension of metaphorical references: while young children capitalize on TOM, older children appear to be boosted by their receptive grammar skills. This pattern is consistent with Deckert et al. (2019) who observed (with another type of metaphor task) that verbal intelligence was associated with better metaphor performance in 9 and 10-year-olds, but not among those who are 7 or 8. They argued that "*early metaphors become gradually more linguistically articulated*" (p. 358).

What accounts for these two interlocking developmental trends? We consider two possibilities. The first is that children simply need to build strong TOM skills to understand metaphors (and potentially other phenomena) before eventually using those skills to later make refined linguistic judgements. Another possibility is that, within the grammar task, items that account for variations among younger children are not the same as those that distinguish abilities among older children. That is, simple items (for example, those targeting the distinction between past and present tenses) would discriminate young children but not older children, who are more likely to show refined abilities when comprehending complex structures (such as embedded clauses). Measures reflecting complex grammatical competence, available to older children, would then avail themselves to tasks that measure metaphor comprehension. These two possibilities are not mutually exclusive.

The apparent developmental shift from TOM-supported metaphor comprehension among children to language-boosted comprehension has theoretical as well as applied implications. Regarding pragmatic theories in general, our results question a dichotomy recently formalized in the literature between linguistic pragmatics and social pragmatics (Andrés-Roqueta & Katsos, 2017, 2020). The first term designates pragmatic cases which depend on structural language and pragmatic norms, while the second relies on those plus competence with TOM. In this framework, scalar implicatures (cases in which weaker terms such as "some" prompt addressees to negate a more informative

[1] Likewise, consider the *Reading the Mind through the Eyes* (RME) task (Baron-Cohen et al., 2001), which asks participants to label mental states based on photographs of eyes; assessments here rely on participants having a vocabulary that distinguishes between descriptions such as *serious, ashamed, alarmed, bewildered* and so on.

term such as “all” so that *some* is often pragmatically understood as *not all*) are typically viewed as a prime example of linguistic pragmatics while irony is considered exemplary of social pragmatics[2]. However, our results with metaphor show that a single pragmatic phenomenon (which these authors would classify as exemplary of linguistic-pragmatics) can be associated with both language and TOM, but at different times of development. While we agree with the authors quest to disentangle the umbrella term “pragmatic inferences” (for a large review see Noveck, 2018), our developmental results indicate that dividing the field up into phenomena is not necessarily the way to do it. At the very least, the current results underline the importance of considering developmental perspectives when building such models.

Despite our optimism that the work here advances discussion about metaphor development, we would be remiss if we did not consider the limitations of the current work. The first is that the present work has considered metaphors as a general phenomenon while not considering specific patterns that might emerge by distinguishing between its subtypes, such as the distinction between sensory as opposed to conceptual metaphors (Van Herwegen et al., 2013), or between physical and mental metaphors, where TOM is likely to have a greater importance for the second type (Lecce et al., 2019). Second, many prior studies used vocabulary measures as an index for language abilities, while we used grammar. This calls for caution when making comparisons among studies that underline linguistic abilities, and for further studies on the specific contribution of different language sub-components. Third, the second study relies on a cross-sectional correlational study; to be more convincing, the developmental trends we observe should be confirmed by longitudinal studies as well (see Del Sette et al., 2020). Finally, we observed *associations* between our variables of interest, rather than directly providing evidence for the causal nature of, and the direction of those relationships. Intervention studies, targeting either TOM of language to improve metaphor comprehension, should now help specifying the nature of those links, while also offering translational perspectives to clinical and educational settings.

To sum up, the current paper adds value to the rich literature on metaphor comprehension development among school-aged children in five critical ways. First, it provides yet another metaphor-reference task to the literature, whose outcomes – like its predecessors – show progressive improvement in metaphor comprehension with age. Second, the current task, unlike prior ones that largely focus on the target reference itself, finely separates reference assignment from a picture selection portion of the task, showing that the metaphor comprehension and slowdown effects reported in the literature are robust, even as a participant’s focus moves downstream (cf. Noveck et al., 2001). Thirdly, to our knowledge, this is the first study to consider the potential association of both structural language and TOM with metaphor comprehension with a single developmental sample. Specifically, the data show that, while young children capitalize on TOM, older children’s improvement seems to take advantage of advancing skills at grammar. Fourthly, the TOM measure employed in the current paper is based on a non-verbal TOM task, making the TOM arguably less biased than those found in previous studies investigating its role in metaphor development. Finally, the task is carried out on a tablet that can be one of dozens carried into a classroom. This allows for congenial, and thus ecologically valid, data collection that engages young participants.

Author note

NP, VB and IN performed conceptualization and methodology, NP performed funding acquisition, investigation and wrote the original draft, NP and LB performed formal analysis, JP provided resources, IN, VB, LB and JP performed writing (review and editing), IN performed supervision. Data and analysis scripts are available on OSF (https://osf.io/myr2c/?view_only=7c8aeebe9837458ea5eb1c365bf20ccd). We would like to thank Julie Penavayre who lent her voice for the audio recordings, the administrators, teachers

[2] The authors do specify that this dichotomy is not meant to apply to pragmatic phenomena *per se* but rather to communicative situations, and to experimental paradigms, in that in some situation scalar implicature might rely on TOM abilities and vice-versa.

and children from Immaculée conception, Saint-Sacrement, and Beynes schools, who gave a precious time for this study, to Caroline Morellet and Isabelle Petit who helped with data collection. This work was supported by grants to the first author from the Hospital Le Vinatier and by the French Ministry of Health (PHRIP-2018-266), and by the XPrag.it 2020 Young Investigator Training Program funded by the ACRI (Associazione di fondazioni e Casse di Risparmio Italiane SpA).

References

Almor, A., Arunachalam, S., & Strickland, B. (2007). When the Creampuff Beat the Boxer: Working Memory, Cost, and Function in Reading Metaphoric Reference. Metaphor and Symbol, 22(2), 169–193. https://doi.org/10.1080/10926480701235478

Andrés-Roqueta, C., & Katsos, N. (2017). The Contribution of Grammar, Vocabulary and Theory of Mind in Pragmatic Language Competence in Children with Autistic Spectrum Disorders. *Frontiers in Psychology*, *8*, 996. https://doi.org/10.3389/fpsyg.2017.00996

Andrés-Roqueta, C., & Katsos, N. (2020). A Distinction Between Linguistic and Social Pragmatics Helps the Precise Characterization of Pragmatic Challenges in Children With Autism Spectrum Disorders and Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research*, *63*(5), 1494–1508. https://doi.org/10.1044/2020_JSLHR-19-00263

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., & Plumb, I. (2001). The “Reading the Mind in the Eyes” Test Revised Version: A Study with Normal Adults, and Adults with Asperger Syndrome or High-functioning Autism. *J. Child Psychol. Psychiat.*, *42*(2), 241–251.

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

Bishop, D. V. M. (2003). Children’s Communication Checklist—2nd Edition. Pearson.

Boyce, W., Torsheim, T., Currie, C., & Zambon, A. (2006). The Family Affluence Scale as a Measure of National Wealth: Validation of an Adolescent Self-Report Measure. *Social Indicators Research*, *78*(3), 473–487. https://doi.org/10.1007/s11205-005-1607-6

Carston, R., & Yan, X. (2022). Metaphorical Referring is Costly. *Book of Abstrats - XPrag 2022*. https://doi.org/10.17605/OSF.IO/C4KP2

Currie, C., Molcho, M., Boyce, W., Holstein, B., Torsheim, T., & Richter, M. (2008). Researching health inequalities in adolescents: The development of the Health Behaviour in School-Aged Children (HBSC) Family Affluence Scale. *Social Science & Medicine*, *66*(6), 1429–1436. https://doi.org/10.1016/j.socscimed.2007.11.024

de Villiers, J. (2007). The interface of language and Theory of Mind. *Lingua*, *117*(11), 1858–1878. https://doi.org/10.1016/j.lingua.2006.11.006

Deckert, M., Schmoeger, M., Schaunig-Busch, I., & Willinger, U. (2019a). Metaphor processing in middle childhood and at the transition to early adolescence: The role of chronological age, mental age, and verbal intelligence. *Journal of Child Language*, *46*(2), 334–367. https://doi.org/10.1017/S0305000918000491

Deckert, M., Schmoeger, M., Schaunig-Busch, I., & Willinger, U. (2019b). Metaphor processing in middle childhood and at the transition to early adolescence: The role of chronological age, mental age, and verbal intelligence. In *Journal of Child Language* (Vol. 46, Issue 2). https://doi.org/10.1017/S0305000918000491

Del Sette, P., Bambini, V., Bischetti, L., & Lecce, S. (2020). Longitudinal associations between theory of mind and metaphor understanding during middle childhood. *Cognitive Development*, *56*, 100958. https://doi.org/10.1016/j.cogdev.2020.100958

Di Sante, M., & Potvin, L. (2022). We Need to Talk About Social Inequalities in Language Development. *American Journal of Speech-Language Pathology*, *31*(4), 1894–1897. https://doi.org/10.1044/2022_AJSLP-21-00326

Gentner, D. (1988). Metaphor as structure mapping: The relational shift. *Child Development*, 47–59.

Gibbs, R. W. (1990). Comprehending figurative referential descriptions. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 56–66. https://doi.org/10.1037/0278-7393.16.1.56

Happé, F. (1993). Communicative competence and theory of mind in autism: A test of relevance theory. *Cognition*, *48*(2), 101–119.

Happé, F. (1994). An advanced test of theory of mind: Understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *Journal of Autism and Developmental Disorders*, *24*(2), 129–154. https://doi.org/10.1007/BF02172093

Heredia, R. R., & Cieślicka, A. B. (Eds.). (2015). *Bilingual Figurative Language Processing*. Cambridge University Press. https://doi.org/10.1017/CBO9781139342100

Kalandadze, T., Bambini, V., & Næss, K.-A. B. (2019). A systematic review and meta-analysis of studies on metaphor comprehension in individuals with autism spectrum disorder: Do task properties matter? *Applied Psycholinguistics*, *40*(6), 1421–1454. https://doi.org/10.1017/S0142716419000328

Kalandadze, T., Norbury, C., Naerland, T., & Naess, K.-A. B. (2018). Figurative Language Comprehension in Individuals with Autism Spectrum Disorder: A Meta-Analytic Review. *Autism: The International Journal of Research and Practice*, *22*(2), 99–117. https://doi.org/10.1177/1362361316668652

Khomsi, A., Khomsi, F., & Pasquet, F. (2007). *BILO - Bilans Informatisés de Langage Oral*. Pearson.

Kogan, N., & Chadrow, M. (1986). Children's comprehension of metaphor in the pictorial and verbal modality. *International Journal of Behavioral Development*, *9*(3), 285–295.

Kogan, N., Connor, K., Gross, A., & Fava, D. (1980). Understanding visual metaphor: Developmental and individual differences. *Monographs of the Society for Research in Child Development*, 1–78.

Langdon, R., & Coltheart, M. (1999). Mentalising, schizotypy, and schizophrenia. *Cognition*, *71*(1), 43–71. https://doi.org/10.1016/S0010-0277(99)00018-9

Lecce, S., Ronchi, L., Del Sette, P., Bischetti, L., & Bambini, V. (2019). Interpreting physical and mental metaphors: Is Theory of Mind associated with pragmatics in middle childhood? *Journal of Child Language*, *46*(2), 393–407. https://doi.org/10.1017/S030500091800048X

Lenth, R. (2022). *emmeans: Estimated Marginal Means, aka Least-Squares Means* ({R package version 1.8.2},). https://CRAN.R-project.org/package=emmeans

Matthews, D., Biney, H., & Abbot-Smith, K. (2018). Individual Differences in Children's Pragmatic Ability: A Review of Associations with Formal Language, Social Cognition, and Executive Functions. *Language Learning and Development*, *14*(3), 186–223. https://doi.org/10.1080/15475441.2018.1455584

Morsanyi, K., Stamenković, D., & Holyoak, K. J. (2020). Metaphor processing in autism: A systematic review and meta-analysis. *Developmental Review*, *57*, 100925. https://doi.org/10.1016/j.dr.2020.100925

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. (2018). *Experimental Pragmatics: The Making of a Cognitive Science*. Cambridge University Press. https://doi.org/10.1017/9781316027073

Noveck, I., 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

Piaget, J. (1964). *Le langage et la pensée*. Éditions Gonthier SA.

Porter, M. A., Coltheart, M., & Langdon, R. (2008). Theory of Mind in Williams Syndrome Assessed Using a Nonverbal Task. *Journal of Autism and Developmental Disorders*, *38*(5), 806–814. https://doi.org/10.1007/s10803-007-0447-4

Pouscoulous, N. (2011). Metaphor: For adults only? *Belgian Journal of Linguistics*, *25*, 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

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Rajkumar, A. P., Yovan, S., Raveendran, A. L., & Russell, P. S. S. (2008). Can only intelligent children do mind reading: The relationship between intelligence and theory of mind in 8 to 11 years old. *Behavioral and Brain Functions*, *4*(1), 51. https://doi.org/10.1186/1744-9081-4-51

Reynolds, R. E., & Schwartz, R. M. (1983). Relation of metaphoric processing to comprehension and memory. *Journal of Educational Psychology*, *75*, 450–459. https://doi.org/10.1037/0022-0663.75.3.450

Rosseel, Y. (2012). **lavaan**: An *R* Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*(2). https://doi.org/10.18637/jss.v048.i02

Royston, P., Altman, D. G., & Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: A bad idea. *Statistics in Medicine*, *25*(1), 127–141. https://doi.org/10.1002/sim.2331

Seigneuric, A., Megherbi, H., Bueno, S., Lebahar, J., & Bianco, M. (2016). Children's comprehension skill and the understanding of nominal metaphors. *Journal of Experimental Child Psychology*, *150*, 346–363. https://doi.org/10.1016/j.jecp.2016.06.008

Sperber, D., & Wilson, D. (1996). *Relevance: Communication and Cognition* (2nd ed.). Wiley-Blackwell.

Tonini, E., Bischetti, L., Del Sette, P., Tosi, E., Lecce, S., & Bambini, V. (2023). On the specificity of the relationship between metaphor skills and Theory of Mind in middle childhood: Reviving the Referential Metaphors task. *Cognition*.

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

Van Rheenen, T. E., & Rossell, S. L. (2013). Picture sequencing task performance indicates theory of mind deficit in bipolar disorder. *Journal of Affective Disorders*, *151*(3), 1132–1134. https://doi.org/10.1016/j.jad.2013.07.009

Vézina, M., Sylvestre, A., & Fossard, M. (2013). Development of a Quebec French Version of the. *Revue canadienne d'orthophonie et d'audiologie*, *37*(2), 156–168.

Vosniadou, S. (1987). Children and metaphors. *Child Development*, 870–885.

Whyte, E. M., & Nelson, K. E. (2015). Trajectories of pragmatic and nonliteral language development in children with autism spectrum disorders. *Journal of Communication Disorders*, *54*, 2–14. https://doi.org/10.1016/j.jcomdis.2015.01.001

Willinger, U., Deckert, M., Schmöger, M., Schaunig-Busch, I., Formann, A. K., & Auff, E. (2019a). Developmental steps in metaphorical language abilities: The influence of age, gender, cognitive flexibility, information processing speed, and analogical reasoning. *Language and Speech*, *62*(2), 207–228.

Willinger, U., Deckert, M., Schmöger, M., Schaunig-Busch, I., Formann, A. K., & Auff, E. (2019b). Developmental Steps in Metaphorical Language Abilities: The Influence of Age, Gender, Cognitive Flexibility, Information Processing Speed, and Analogical Reasoning. *Language and Speech*, *62*(2), 207–228. https://doi.org/10.1177/0023830917746552

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

Supplementary Materials

Experiment 1

Table S1.1
Vignette characteristics

	Metaphor-base	Synonymic-base	Stat. comparison
Mean length in words (SD)	54.1 (4.0)	54.0 (4.1)	Z = -0.2, p > .05
Mean readability (SD)[1]	52.4 (8.6)	53.0 (6.8)	Z = -0.4, p > .05
Mean content words frequency (SD)[2]	5.7 (0.5)	5.3 (0.4)	Z = -1.3, p > .05
Mean referent frequency (SD)[3]	59.1 (7.0)	56.9 (9.2)	Z = -0.5, p > .05
Mean target noun frequency (SD)[3]	51.5 (10.3)		/

[1] As indexed by the French tool Scolarius (www.scolarius.com); a high score indicates low readability and scores ranging from 50 to 89 are considered suitable for primary school pupils.
[2] As indexed by the French movie-subtitles database Lexique 2 (New et al., 2004); the median frequency was high, at 297 per million occurrences; this value is at the 97th percentile of the entire database. Since frequencies were not distributed normally, we used the log-transformed data to assess difference across conditions.
[3] As measured in the Manulex database (Lété, 2003) which gathers French Textbooks for primary schools (1st to 5th grade). The overall interquartile range in this database is [24-46].

Table S1.2
By-item results

	Prop. of metaphoric interpretations in the whole sample		Prop. of metaphoric interpretations in the metaphor condition		Mean log-transformed RT (SD)	
Item	Literal context	Metaphor context	7 years-old	10 years-old	Literal context	Metaphoric context
1	0%	46%	18%	71%	7,73 (0,48)	8,01 (0,42)
2	9%	36%	18%	55%	7,92 (0,40)	8,09 (0,38)
3	0%	42%	23%	71%	7,85 (0,44)	7,87 (0,42)
4	11%	36%	18%	55%	7,72 (0,50)	8,10 (0,39)
5	2%	42%	27%	65%	8,07 (0,36)	8,36 (0,29)
6	0%	38%	27%	52%	7,58 (0,34)	7,74 (0,57)
7	4%	19%	5%	35%	8,02 (0,35)	8,37 (0,36)
8	2%	33%	9%	54%	7,78 (0,45)	8,23 (0,39)
9	2%	57%	36%	75%	7,82 (0,46)	8,08 (0,39)
10	2%	32%	12%	57%	7,99 (0,36)	8,04 (0,46)
11	0%	57%	42%	76%	7,94 (0,27)	7,86 (0,44)
12	0%	72%	58%	94%	7,92 (0,43)	8,01 (0,43)

Figure S1.1

Responses provided by children from the 2 groups to the 2 conditions, with the proportion of correct answers

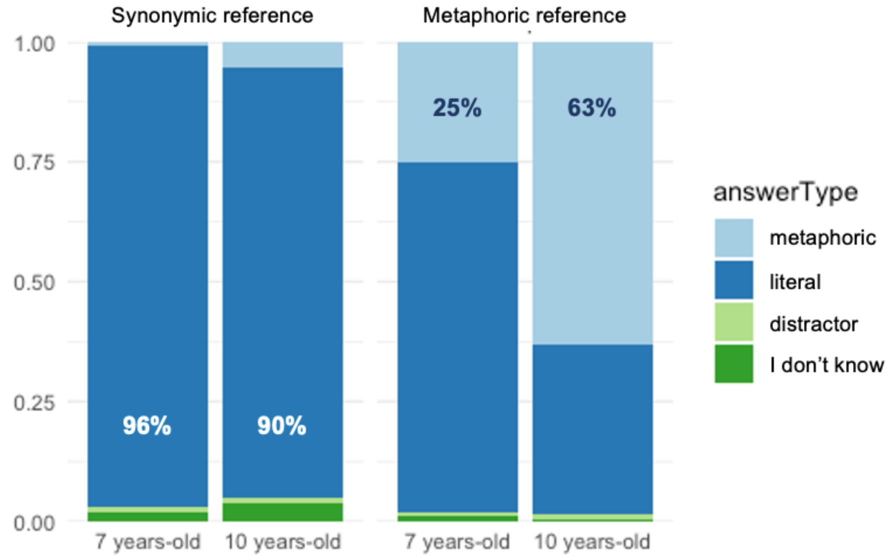


Figure S1.2

Condition and group effects on response times

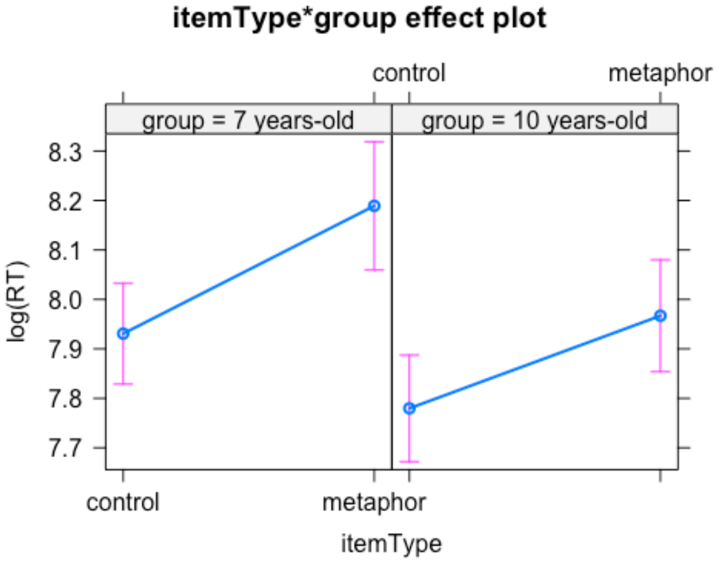


Figure S1.3

List effects on response times, in interaction with condition (left) and group (right)

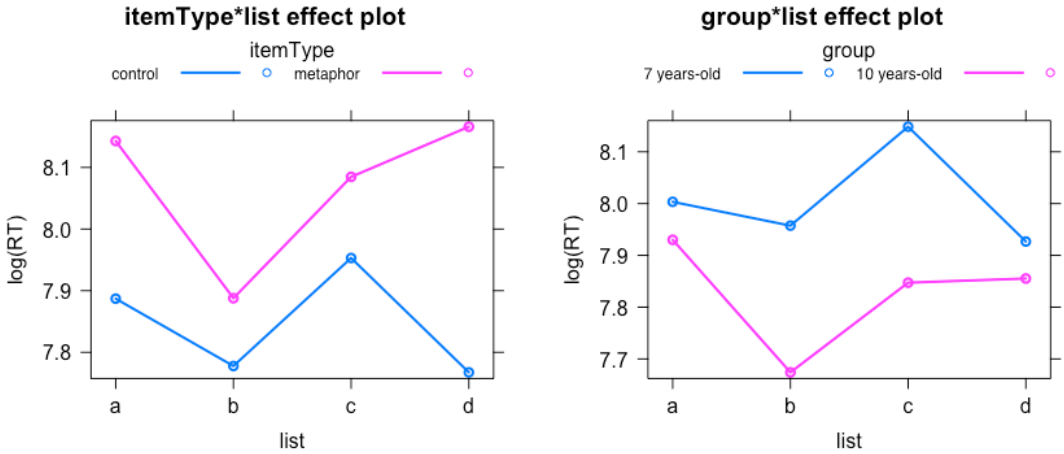
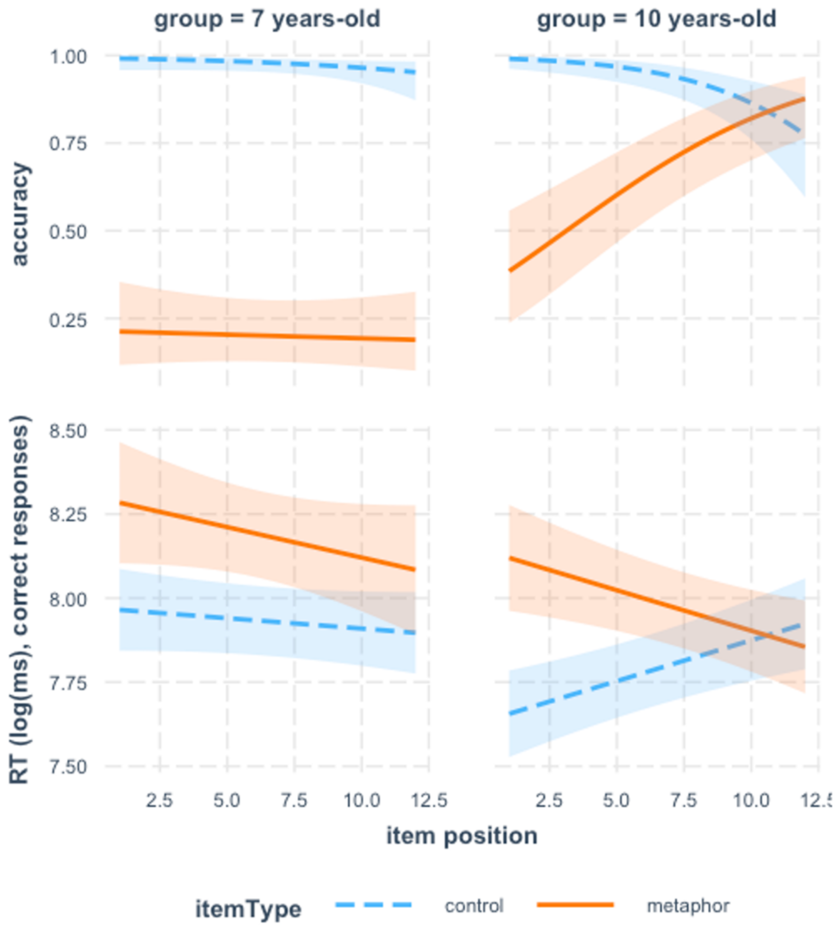


Figure S1.4

Item position effect on accuracy (up) and response times (down) across group and condition



Experiment 2

Table S2.1

The characteristics of Experiment 2's stories across its two conditions

	Metaphor-base	Synonymic-base	Stat. comparison
Mean length in words (SD)	55.3 (3.9)	54.3 (3.6)	$U = 18, p > .05$
Mean readability (SD)[1]	52.2 (5.9)	49.0 (5.1)	$U = 12.5, p > .05$
Mean content words frequency (SD)[2]	5.7 (0.5)	5.6 (0.6)	$U = 16, p > .05$
Mean referent frequency (SD)[3]	56.5 (7.6)	52.4 (10.7)	$U = 13, p > .05$
Mean target noun frequency (SD)[3]	55.1 (6.3)	48.0 (12.7)	$U = 9, p > .05$

[1] As indexed by the French tool Scolarius (www.scolarius.com); high score indicate low readability and scores ranging from 50 to 89 being suitable for primary school pupils.

[2] As indexed by the French movie-subtitles database Lexique 2 (New et al., 2004). The median frequency was high, at 297 per million occurrences, this value being at the 97th percentile of the whole database. Since frequencies were not distributed normally, we used the log-transformed data to assess difference across conditions.

[3] As measured in the Manulex database (Lété, 2003) gathering French Textbooks for primary schools (1st to 5th grade). Mean frequency was 54.41 (SD = 9.1) for referents and 51.4 (SD = 10.3) for targets, the overall interquartile range in this database being [24-46].

Table S2.2

Estimated TOM effect on accuracy in the different conditions, at different ages, with p-value of the pairwise difference at each age point

Age point	Condition	TOM estimated trend	SE	*p* value for condition difference
6 years old	Synonymic	-0.33	0.24	< .01
	Metaphoric	0.34	0.16	
7 years old	Synonymic	-0.24	0.18	< .01
	Metaphoric	0.25	0.11	
8 years old	Synonymic	-0.15	0.15	< .05
	Metaphoric	0.15	0.09	
9 years old	Synonymic	-0.06	0.17	.50
	Metaphoric	0.05	0.12	
10 years old	Synonymic	0.03	0.24	.76
	Metaphoric	-0.04	0.17	

Table S2.3

Estimated Grammar effect on accuracy in the different conditions, at different ages, with p-value of the pairwise difference at each age point

Age point	Condition	Grammar estimated trend	SE	*p* value for condition difference
6 years old	Synonymic	0.08	0.07	< .05
	Metaphoric	-0.08	0.05	
7 years old	Synonymic	0.02	0.05	.99
	Metaphoric	0.02	0.04	
8 years old	Synonymic	-0.04	0.06	< .05
	Metaphoric	0.12	0.04	
9 years old	Synonymic	-0.09	0.07	< .001
	Metaphoric	0.21	0.05	
10 years old	Synonymic	-0.15	0.10	< .001
	Metaphoric	0.31	0.06	

Table S2.4

Estimated TOM effect on log-transformed response times in the different conditions, at different ages, with p-value of the pairwise difference at each age point

Age point	Condition	TOM estimated trend	SE	*p* value for condition difference
6 years old	Synonymic	-0.01	0.02	< .05
	Metaphoric	-0.10	0.03	
7 years old	Synonymic	-0.01	0.02	< .05
	Metaphoric	-0.07	0.03	
8 years old	Synonymic	-0.01	0.01	.13
	Metaphoric	-0.04	0.02	
9 years old	Synonymic	-0.005	0.04	.82
	Metaphoric	-0.001	0.02	
10 years old	Synonymic	-0.002	0.03	.20
	Metaphoric	-0.03	0.03	

Table S2.5

Estimated Grammar effect on log-transformed response times in the different conditions, at different ages, with p-value of the pairwise difference at each age point

Age point	Condition	Grammar estimated trend	SE	*p* value for condition difference
6 years old	Synonymic	-0.07	0.008	.41
	Metaphoric	0.002	0.01	
7 years old	Synonymic	0.0002	0.006	.98
	Metaphoric	0.0004	0.008	
8 years old	Synonymic	0.007	0.007	.17
	Metaphoric	-0.002	0.008	
9 years old	Synonymic	0.01	0.007	< .05
	Metaphoric	-0.004	0.008	
10 years old	Synonymic	0.02	0.003	< .01
	Metaphoric	-0.005	0.01	

Table S2.6

Estimated TOM effect on accuracy in the different conditions, at different ages, with p-value of the pairwise difference at each age point

Age point	Contrast	Ratio	SE	*t*-ratio	*p* value
6 years old	**Metaphor-lit vs. metaphor-pragma**	**1.28**	**0.064**	**-0.25**	**< .0001**
	Metaphor-lit vs. synonymic	0.99	0.058	5.02	0.97
7 years old	**Metaphor-lit vs. metaphor-pragma**	**1.20**	**0.036**	**6.08**	**< .0001**
	Metaphor-lit vs. synonymic	1.02	0.056	0.30	0.95
8 years old	**Metaphor-lit vs. metaphor-pragma**	**1.16**	**0.039**	**4.39**	**< .0001**
	Metaphor-lit vs. synonymic	1.04	0.060	0.75	0.74
9 years old	**Metaphor-lit vs. metaphor-pragma**	**1.16**	**0.036**	**4.71**	**< .0001**
	Metaphor-lit vs. synonymic	1.07	0.061	1.13	0.51
10 years old	**Metaphor-lit vs. metaphor-pragma**	**1.19**	**0.057**	**3.66**	**< .001**
	Metaphor-lit vs. synonymic	1.08	0.071	1.23	0.45

Figure S2.1

Correlation matrix of all predictors in Experiment 2

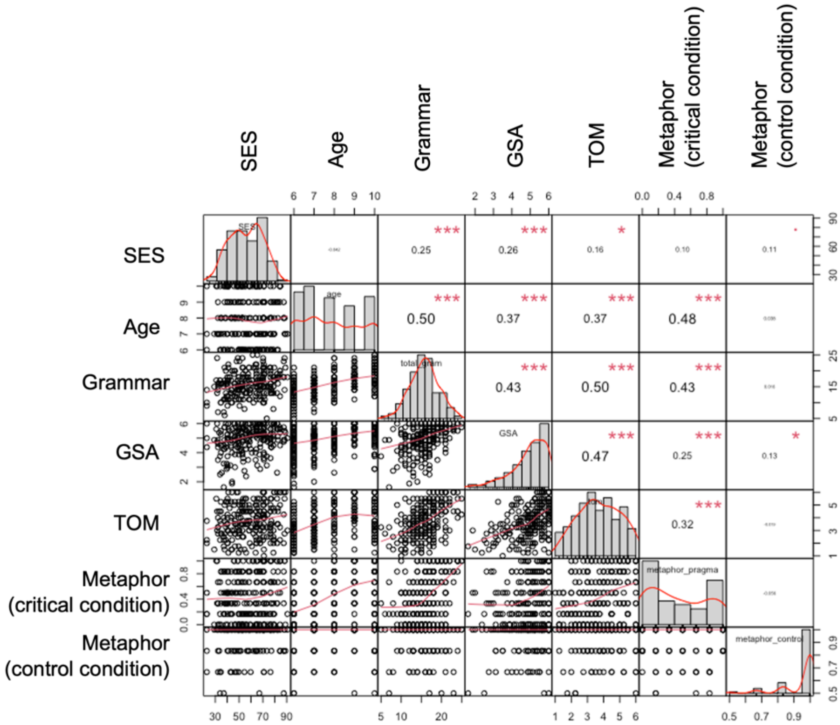


Figure S2.2

Log-transformed response times distribution (plain line) against a normal distribution with the same parameters (dashed line)

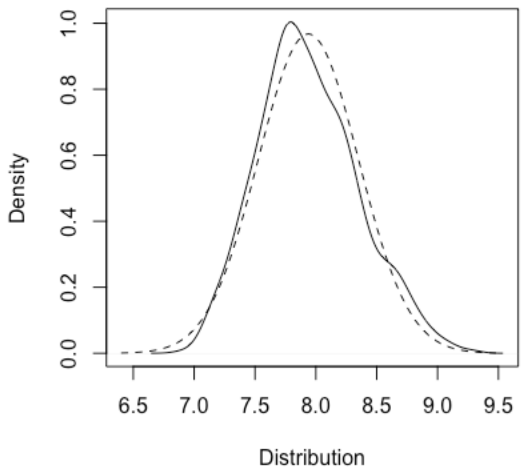


Figure S2.3

Exploratory analysis: predicted response times as a function of age for correct responses to literal items, correct (metaphorical) responses to metaphor items and literal responses to metaphor items