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

Nicolas Petit, Valentina Bambini, Luca Bischetti, Jérôme Prado, Ira Noveck

How do theory of mind and formal 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 Psychiatrie Universitaire (Lyon Métropole, FRANCE)
[2] Centre de Recherche en Neurosciences de Lyon, INSERM U1028 - CNRS UMR5292 - Université 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, CNRS-Université Paris Cité (Paris, FRANCE)

* Corresponding author, e-mail: npetit.ortho@gmail.com

This manuscript was accepted for publication in the Journal of Experimental Psychology: Learning, Memory, and Cognition. The DOI of the final publication is : https://doi.org/10.1037/xlm0001381

Author Note

Nicolas Petit https://orcid.org/**0000-0002-9910-2706**

Conceptualization and methodolgy: NP, VB, IN; Funding acquisition, investigation and writing (initial draft): NP; Ressources: JP; Formal analysis: NP, LB; Writing (review and editing): IN, VB, LB, JP; Supervision : IN. Data and analysis scripts are available on OSF (Petit, 2024, https://osf.io/myr2c/).

We would like to thank Julie Penavayre, who lent her voice for the audio recordings, as well as to the administrators, teachers and children from Immaculée conception, Saint-Sacrement, and Beynes schools, who gave their precious time so that we could carry out this study. Likewise, we thank 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).

Abstract

Prior research has shown that school-aged children's metaphor comprehension becomes adult-like progressively. This has given rise to claims that the development of metaphor comprehension is due to children's evolving abilities with respect to Theory of Mind (ToM) or to formal language. The present work investigates the extent to which children's growing sophistication with metaphor is attributable to each of these. Experiment 1 validates a newly constructed tablet task – with two groups of children whose mean ages were approximately 7 and 10 (N=89) – in which participants a) listen to vignettes that conclude with either a metaphoric or a synonymic (control) reference and then; b) choose pictures (while latencies are recorded) that indicate whether the children understand the reference as intended. The outcomes from Experiment 1 confirm prior results: Accurate

responding in the wake of a metaphoric reference increases with age; meanwhile, correct metaphoric responses take longer than synonymic ones. Experiment 2 tests a more expansive range of 6- to 11-year-olds (N=248) and a wider array of tasks, including two clinical tasks measuring ToM and formal language skills which we use as cognitive predictors of metaphor accuracy and response times. Results show that ToM is a reliable predictor of successful performance on the metaphor task among younger children, before attenuating with age; in contrast, formal language is a predictor of metaphor comprehension that strengthens with age and is maximal in older children. This work underlines the importance of considering developmental perspectives when investigating the cognitive bases of metaphor skills.

Keywords: metaphor, pragmatic development, theory of mind, reference, language development

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

Introduction

Children's metaphor comprehension becomes adult-like with age. This is revealed by tasks that require young participants to act out a metaphoric description, to satisfactorily explain a metaphor (Evans & Gamble, 1988), or to naturally read lines of text that incorporate a metaphoric reference (Noveck et al., 2001). Children's school-age years appear especially critical to this development. A natural question that follows from the results of such studies is what are the underlying or associated cognitive abilities that support this development? Based on adjacent work inspired by investigations into atypical development (e.g. Happé, 1993), one proposal is that children's improvement comes from their increasingly refined ability to understand a speaker's *intended* meaning, which points to abilities related to Theory of Mind (ToM). Another line of research, also influenced by studies on atypical populations, is that formal language skills themselves continually develop and that these naturally play a role in the development of figurative language comprehension (e.g. Norbury, 2005).

In the remainder of this Introduction, we take the following three steps. First, we summarize in greater detail the studies that investigate school-aged children and their growing abilities with metaphor. Second, we briefly review findings that aim to uncover the cognitive abilities that support metaphor comprehension. This review will reveal how Theory of Mind (ToM) and formal language skills are the leading candidates for describing what is at the core of children's growing metaphor comprehension. Finally, we will introduce our two Experiments, which investigate children's accuracy and speed as they identify metaphoric references, as well as their synonymic controls, while listening to orally presented vignettes via a tablet.

The development of metaphoric comprehension among school-aged children

There is a rich history behind investigations of metaphor development (to appreciate the early studies, see a review from Vosniadou, 1987) but in order to present some relevant background for the current work we start by considering Winner et al. (1976), who concluded that children need to be about 14 before fully appreciating the intended meaning of metaphorical sentences, such as *The prison guard was a hard rock*. This kind of claim was consistent with Piagetian views which assumed that metaphor comprehension was late developing. Nevertheless, researchers soon pointed out that such a claim does not mean that children's metaphor comprehension abilities are non-existent until they are older. Both classical and modern cognitive studies have shown that metaphor comprehension is an *evolving* ability. This becomes apparent through experimental manipulations that facilitate children's understanding of metaphor. For example, Vosniadou et al. (1984) showed that very young participants' metaphor comprehension is more adult-like when a critical test expression is part of a probable, as opposed to an improbable, ending. Similarly, Lecce et al. (2019) showed that nine-year-olds are more competent at describing "physical" metaphors, such as *Dancers are butterflies*, than they are at describing "mental" metaphors, such as *Soldiers are lions*. These are among the clues that indicate that school-aged children's metaphor comprehension is a work in progress. Below, we review two lines of research whose robust findings generally support the claim that mastery over (i.e., detecting and comprehending) metaphor evolves during children's school-age years.

Triad tasks

Consider an original task (first introduced by Kogan et al. [1980]) in which children are put into a position to *detect* and *explain* metaphoric relations among three items in what is called the *Metaphor Triad Task* (the MTT). Here, young participants are presented with three pictures (or words) that allow up to three pairing possibilities, one of which is designed to be metaphoric in character. For example, consider participants who are presented pictures of a *fish, a winding river,* and *a snake* while noting that the latter two would be expected to be the source of a metaphoric relation. For each triad, participants select what they consider to be the best possible pairing and are asked to explain it. They are then asked to consider the remaining two pairs (in this way the metaphoric pairing, if initially bypassed, is at some point addressed). Findings from this kind of task underscore how developmental patterns emerge.

Using this paradigm, Deckert et al. (2019) tested second to fourth graders (i.e., children between the ages of 7 and 11) and recorded (a) whether or not the metaphorical relation was identified at all (*identification*); if so, (b) whether or not it was correctly explained (*explanation*) and finally; (c) whether or not the metaphoric relation was identified first (*preference*). The results show that there are at least two identifiable cognitive leaps. One occurs at around 8 years of age, at which point children begin to *identify* the metaphoric pairs and *explain* them better. The second one is observable at around 10 years of age, at which point participants increasingly *prefer* metaphoric pairs over non-metaphoric ones (also see, Willinger et al., 2019). These shifts provide the literature with relevant markers concerning children's metaphoric-language maturation.

Reading comprehension tasks

A second research stream comes from studies that use vignettes that include metaphoric references. Modeled on Gibbs's (1990) study with adults, Noveck et al. (2001) presented 8- to 12-year-old French-speaking children with a reading task whose penultimate sentence referred back to a feature in the text that had been mentioned earlier. For example, one of their stories described a class of second-grade pupils (*éleves*) who were having a swimming lesson in a pool and who are later referred to, by the instructor in the story, as either "toads" (*crapauds*) metaphorically (in one condition) or else as "students" (*étudiants*) synonymically (in the other). 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, at which point the gap between the two conditions appears to close. In a second experiment, they presented the same stories in the form of a self-paced reading task to children who were 8, 11, and 14, as well as to adults in order to capture latencies for the referential sentences. Their data revealed – for children as well as for adults – that metaphoric references took reliably longer to read than their synonymic controls; this indicates that there is a cost in making metaphoric references across all ages. Like in their Experiment 1, rates of correct responses to the comprehension questions again showed that an initial disadvantage among the youngest children for post-metaphoric questions (compared to post-synonymic ones) diminishes with age.

Similar referential tasks have been used in other developmental studies (Van Herwegen et al., 2013; Seigneuric et al., 2016; Tonini et al., 2023). One of these (Van Herwegen et al., 2013) presented its stories orally while also using pictures (so that young participants and those with reading

difficulties could be readily included); it too has shown that metaphor comprehension increases significantly with age among typically developing children. This innovation (using pictures along with an oral presentation) will become relevant to the work here. We also add that other adult studies have confirmed that there is a temporal cost associated with metaphor processing when compared to yoked controls (Almor et al., 2007; Gibbs, 1990; Heredia & Cieślicka, 2015; though for nuance see Carston & Yan, 2023).

Substrates of metaphor comprehension among children

Here, we turn to the underlying substrates that arguably play a role in children's growing metaphorical skills (for a review, see Kalandadze et al., 2019). As we indicated earlier, one notable candidate is Theory of Mind (ToM). This possibility emerged because it has long been noted that autistic individuals tend to prefer more literal readings of metaphoric expressions and, as a group, they are known to have difficulties with intention reading (or ToM). It was Happé (1993) who first made the link between autistic children's understanding of metaphor and irony and their level of theory of mind ability. Happé (1993) employed Relevance Theory (Sperber & Wilson, 1996), a Gricean inspired cognitive account focused on intended readings and communication, in order to address the way a metaphor's meaning is understood via inferences towards the speaker's intention. Several studies have followed up on this insight.

One comes from Lecce et al. (2019), whose work was briefly mentioned earlier. They evaluated children's performance in explaining physical and mental metaphors while using sentences having the form "X is Y." This work was based, in part, on an analysis showing that older (10- and 11-year-old) children rely on psychological features more than the youngest (9-year-old) children in the study. The authors also recorded measures of ToM abilities from each of the children, through the Strange Stories task (Happé, 1994). Interestingly, they observed that, when this ToM measure was used as a statistical predictor of outcomes, it was associated with psychological (but not physical) metaphor interpretation among children at 9 years of age, even after controlling for vocabulary, SES and working memory. This was not the case for participants who were 10 or 11 years of age. They thus proposed that ToM's influence on metaphor interpretation is early or short-lived as children grow (see also Tonini et al., 2023 for similar evidence). This result resonated with findings from a short-term longitudinal study from the same team, in which Del Sette et al. (2021) observed an association between metaphoric explanations and Theory of Mind growth among 9-year-olds.

As we indicated earlier, Happé's seminal work led to an alternative proposal from Norbury (2005), who argued that the development of *formal language skills* is behind improvement of metaphor comprehension abilities. According to this proposal, abilities with respect to semantics would be an appropriate predictor of children's metaphor comprehension because a skilled figurative comprehender needs to appreciate the relevant traits of the vehicle e.g., the word *shark* in *my lawyer is a shark* as it is joined with the topic (*lawyer* in this example). That is, Norbury astutely pointed out that understanding the intended meaning of the metaphor *My lawyer is a shark* compels the listener to ignore the fact that sharks are fish with fins and so on while recognizing that other features – that sharks are aggressive and vicious – remain relevant. Furthermore, the traits that make for successful metaphors are often not the vehicle's most salient ones. This implies that having relevant world knowledge and adequate sophisticated semantic representations are required when discerning metaphors (for adult work in this line, see Rubio-Fernandez [2007]).

One can find supporting elements for the *formal language skills* claim in the existing literature. Deckert et al. (2019) analyzed children's performance on a version of the Metaphor Triad Task and reported that scores of verbal intelligence and linguistic competence did not predict seven- and eight-year-olds' performance but that verbal intelligence scores did for the nine- and ten-year-olds. In a French study, Seigneuric et al. (2016) compared two groups of eight- to ten-year-old children that differed in text comprehension abilities (but matched for vocabulary and decoding) as they made literal and metaphoric references. Whereas "poor comprehenders" were found to be statistically weaker than "good comprehenders" in identifying literal references (e.g., in recognizing that "snake" refers back to "a viper"), an interaction revealed that the "poor comprehenders" were especially challenged in identifying metaphoric references (e.g., in recognizing that "butterflies" refers back to "the dancers").

These two abilities – one centered on Theory of Mind and the other on formal language skills – have emerged as two of the most likely candidates for providing the cognitive scaffolding of metaphor comprehension among autistic individuals (for a review, see Kalandadze et al., 2019). It makes sense then to consider these two substrates as sources of neurotypical development of metaphor comprehension as well. It was in this vein that Whyte & Nelson (2015) tested typically developing children between the ages of 5 and 12 on the *Comprehensive Assessment of Spoken Language* (CASL), which among its subtasks are the comprehension of metaphor as well as sarcasm and indirect requests. They observed that both ToM, as 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 rates of non-literal 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 two issues call for further attention. One is that it is not clear whether the two abilities work in tandem or separately. The other is that Whyte and Nelson's study did not single out children's performance on the comprehension of metaphor. As far as we know, no study has aimed to disentangle the relative influence of ToM and formal language skills on metaphor comprehension in a single developmental study. Lecce et al. (2019) and Tonini et al. (2023) did acknowledge the potential influence of these two factors, by controlling for language skills when they assessed ToM's influence on metaphors. However, their general focus was on ToM as they reported their results and conclusions.

Taken together, prior work points to the existence of underlying cognitive abilities that influence metaphor development. Our specific question is *whether* the two abilities we consider here – ToM and formal language abilities – manifest themselves as critical to metaphor comprehension in child development and, if they do, *when*. One possibility, based on prior findings, is that these supporting factors do *not* emerge together in order to support metaphor comprehension in a child's development. In fact, the existing data point to the idea that an individual factor, whether it be ToM or language skills, does not lead to linear growth over children's school age years. It is thus conceivable that a specific competency has a passing influence on the development of metaphor comprehension. In other words, it is possible that ToM and formal language skills influence metaphor comprehension development at specific points in a child's cognitive life.

Goals of the current study

The current work has two general goals. One is to introduce a newly constructed self-paced metaphoric reference task whose developmental outcomes are expected to be consistent with prior

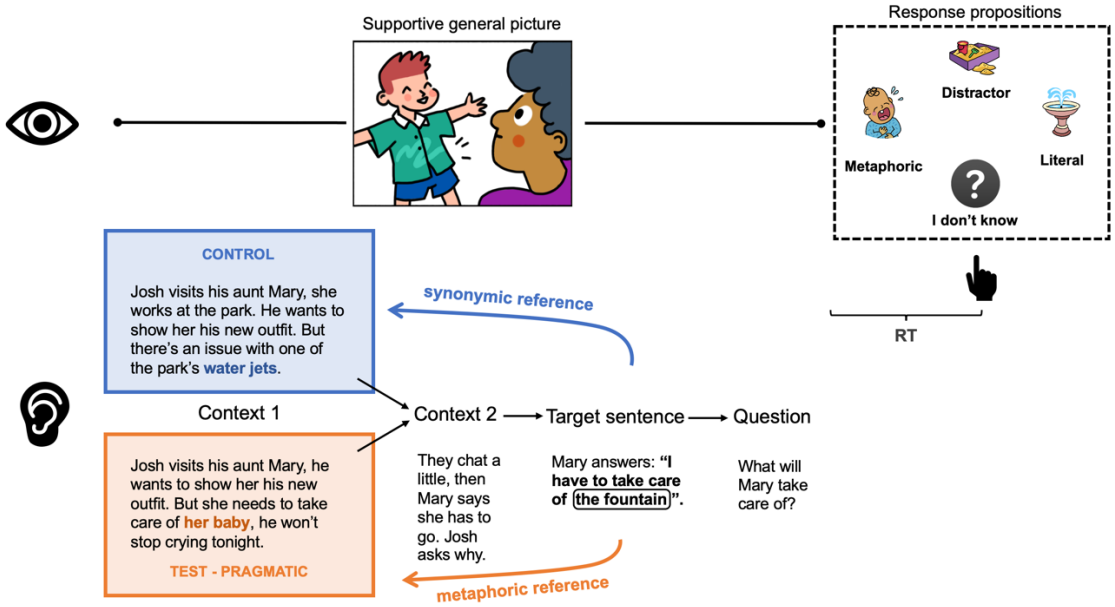
studies (e.g., Noveck et al., 2001; Seigneuric et al., 2016; Van Herwegen et al., 2013). The other is to have participants further provide measures of ToM and formal language abilities so that we can determine what factors influence metaphor comprehension success during children's school-age years. These two goals will be addressed in two separate experiments. In the first, we simply validate the task. In the second, we administer the same task under conditions resembling a cognitive assessment, so that we can carefully investigate children's metaphor comprehension development while also collecting measures regarding their ToM and formal language skills.

Experiment 1

Experiment 1 aims to validate our metaphoric reference tablet task, which consists of a series of vignettes. Each of the vignettes is presented aurally and is designed to introduce a narrative element that will later be referred to in its concluding target sentence (see **Figure 1**). Crucially, the sentence that contains the reference is designed to be metaphorical or else synonymic. At the end of each vignette, participants are presented with a question, which essentially asks them to identify the reference through a picture selection that includes four options. Importantly, the intended target in the metaphor condition, is presented alongside a foil, which is a pictorial representation of the reference understood literally. The children's choices and the speed at which they are provided are recorded.

Figure 1

An illustration of a yoked pair of items (translated from French to English), depicting what a participant sees (top) and hears (bottom) as a vignette unfolds.



Here, we point out one innovative aspect of the experiment, which is that the task is presented by way of a tablet. That is, each child – in a classroom – carries out the task on a provided tablet while listening to instructions and items via headphones. This allows young participants to work at their own pace in the comfort of their classroom. Based on our experience, this approach also has the advantage of being engaging for children.

Given that Experiment 1 aims to validate the task, it is not overly concerned with determining specific developmental milestones but rather in confirming the developmental sensitivity of the task. Namely, we expect older school-aged children to show more competence at metaphoric reference than younger ones. We also expect that the speed at which participants choose the correct picture to be generally slower in the metaphoric condition than in the synonymic condition. We begin by comparing two groups of children that are representative of younger and older school-aged children: 7-year-olds and 10-year-olds.

Method

Transparency and openness

All data and analysis code are available at https://osf.io/myr2c/ (Petit, 2024). This experiment design and its analysis were not pre-registered.

Participants

Eighty-nine native French-speaking children were recruited from a local private school. Parents reported that the children had no neurodevelopmental disorders (including those related to learning, autism spectrum, attention-deficit, or language) nor any motor or sensory (i.e., visual or auditory) disorders that would prevent their children from using a tablet. The study placed children into two age groups: young children from 1st and 2nd grade (N = 48, 23 girls, mean age = 7 years and 0 months [7;0], SD = 0;7) and older children from 4th and 5th grade (N = 41, 13 girls, mean age = 10;0, SD = 0;7). Ethnicity is not reported since the collection of such information is forbidden by French law. 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 is a very short story, or *vignette*, that is orally presented in French. It is voiced by the first author and divided into three parts (see **Figure 1**). The first part (see *Context 1* in Figure 1) consists of two to three sentences that provide relevant background information; this includes mention of a noun phrase that differs as a function of Experimental condition. In one case, it lays the groundwork for a metaphoric reference (e.g., by mentioning a *crying baby* in Figure 1), while in the other, it lays the groundwork for a synonymic reference (e.g., by mentioning *water jets)*. The second part, Context 2, is a sentence that is found in both conditions. Finally, there is a *Target sentence* that is either metaphoric or synonymic in light of the prior context. Returning to **Figure 1**, the reference *fountain* can be used to refer back to the *crying baby* in what will be called the Metaphoric Reference condition or to the *water jets* in what will be called the Synonymic Reference condition.[1] Importantly, all metaphoric target expressions were designed to be novel. Another

[1] Note that this is unlike Noveck et al.'s task (2001), which used different (metaphoric or synonymic) target words as part of an utterance in order to refer back to a singular previously-mentioned element in the provided context.

general feature of the paradigm is that each vignette is accompanied by a picture that is thematically related to it in order to keep the children engaged as they are listening.

The task is self-paced. That is, in order to advance in the vignette, participants are required to tap a speaker button (and they can listen to a segment as often as they like through a 'repeat' button). After the presentation of the vignette, 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 question is completed, four pictorial options are displayed on the screen. From the point of view of the metaphoric condition in **Figure 1**, the picture options reflect a) a representation of the metaphoric reference that harks back to the intended referent (e.g. Figure 1's 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" (IDK) symbol. Both the participants' picture selection and their selection response times (RT) are automatically recorded by the tablet.

From here on, 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 (e.g. the *crying baby*). On the other hand, after a vignette makes a *synonymic* reference, an accurate response involves choosing the picture that corresponds to the reference's literal meaning. 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, we expect accurate responses in the Synonymic Reference condition to be uniformly high across both age groups.

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

Experiment 1 was designed so that a participant received one version of each yoked pair of vignettes (i.e., one of its two versions) and equal amounts of metaphoric and synonymic versions. We prepared four lists of items, each consisting of six vignettes that ultimately present a metaphoric reference and six that present a synonymic one. All told, each vignette theme appeared in its metaphoric version in two lists and in its synonymic version in the two other lists, so that each version of each vignette appeared in two different lists of items. The presentation order of the 12 vignettes within each of the four lists (once they were rendered metaphoric or synonymic) was based on a randomized procedure. In order to control for potential trial-order effects each list order was also reversed so that we ultimately employed 8 different lists.

To familiarize participants with the task and its 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 anywhere from 15 to 25 children completing the task at a time. Each child was provided with a tablet and headphones. Before beginning the test session, each piece of equipment was systematically checked, along with the accompanying app. Each tablet had one of the eight lists randomly assigned to it.

The children, as a group, were told 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.

During preparations in the classroom, the children were told 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 task were analyzed with (generalized or linear) 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 and after removing outliers, which were cases in which response times were at least ±2 *SD* from each age group's mean, within each condition. Sum contrasts were used for all independent variables.

Results

Before beginning our developmental analyses, we first verified that the lists, which were randomly attributed to children, were evenly distributed across age groups; as expected, the children allocated to each list did not differ in age ($\chi^2(3) = 0.5$, $p = .66$).

Accuracy

Overall, accuracy in the Synonymic Reference condition was very high (93%), indicating that participants correctly chose the intended target. As expected, rates of correct picture choices in the Metaphoric Reference condition were lower (43%). We consider each of these findings in greater detail below.

We constructed a generalized model using Age group (younger children, older children) and Reference condition (Metaphoric, Synonymic) as fixed effects, as well as their interaction, with random intercepts for items and participants. This revealed a main effect of Reference condition (ß = 3.4, *SE* = 0.25, *z* = 14.0, $p < .001$) and an interaction of Reference condition by Age group (ß = 3.2, *SE* = 0.45, *z* = -6.9, $p < .001$). As far as the metaphoric reference task is concerned, we observed that participants' picture-choices provided a clear developmental pattern. Twenty-five percent of the younger children's choices in the metaphoric condition were accurate, while 63% of the older children's choices were (ß = -2.0, *SE* = 0.28, *z* ratio = 7.3, $p < .001$). As anticipated, nearly all the errors (97%) among the metaphoric items were due to children's choosing pictures that depict the literal

representation of the target reference (see **Figure S1.1** in supplementary materials). The two other response options were very rarely chosen.

Interestingly, the contrast within Age group concerning the Synonymic Reference condition was significant but in a direction opposite of what one might expect. While practically at ceiling, the younger children committed slightly fewer errors (4%) than the older children (10%; ß = 1.2, *SE* = 0.4, *z* ratio = 2.7, *p* = .006). This result appears to indicate that the older children become wary of alternative readings which extended to the synonymic condition. This trend re-emerges elsewhere in the results and will also be addressed in the Discussion.

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 and considered 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 the supplementary materials). These 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. Moreover, after a metaphoric reference, older children were more likely than younger children 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 *p*s > .05), but clear effects in the older group. That is, older children's 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 Reference condition (ß = -0.31, *SE* = 0.08, *z* ratio = -3.8, *p* < .001). This is in line with the finding concerning the older children's performance in the Synonymic Reference condition discussed earlier. We also added gender as a fixed effect to the initial model (as well as its interaction with the two other fixed effects). This did not improve the model's fit ($\chi^2(4) = 3.1$ $p = .55$), confirming that the results remained very similar across gender.

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 mixed effect model based on the same structure as the one we used for accuracy in which Age group (younger children, older children) and Reference condition (Metaphoric, Synonymic) are fixed effects, along with their interaction, and with random intercepts for items and participants.

This model revealed a main effect of Age group (ß = 0.19, *SE* = 0.06, *t* = 3.3, *p* = .001), 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 with a slowdown due to the Metaphoric condition (ß = -0.22, *SE* = 0.03, *t* = -7.0 *p* < .001), and there was no interaction (ß = -0.07, *SE* = 0.06, *t* = -1.1, *p* = .26). Adding *list* as a fixed factor improved the model's fit ($\chi^2(8) = 21.1$, $p = .007$). While certain lists appeared more sensitive to effects of Reference condition 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** among the supplementary materials).

Finally, we added *item position* (*viz* trial number) 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 younger children (all *p*s > .05). In contrast, the older children tended to respond faster in correctly choosing their picture in the Metaphoric Reference condition as the task progressed (ß = -0.02, *SE* = 0.008, *t* = -2.8, *p* < .01). The older children also tended to respond more *slowly* in correctly choosing their picture in the Synonymic Reference condition as the task progressed (ß = 0.02, *SE* = 0.007, *t* = 3.5, *p* < .001).

Discussion

In Experiment 1, we tested a new tablet-based metaphoric 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 practicable for children while producing results that resonate with prior findings. These aims were achieved.

Based on the results of the current Experiment, we can draw the following five conclusions. First, given children's performance with the control items (in the Synonymic Reference condition), one can confidently conclude that all of the younger participants are carrying out the task competently. Error rates are low overall for both age groups. Second, the data show that children are sensitive to a vignette's preceding context, providing evidence for the task's face validity. Vignettes that call for a metaphoric reference prompt more equivocality (in their picture choices) among all the children. Third, and most importantly to our study, rates of accurate (intended) metaphoric choices increase with age: the older children are more likely than the younger children to choose pictures that indicate that they made metaphoric references. This is consistent with findings from prior studies (e.g. Noveck, 2001; Seigneuric et al., 2016; Van Herwegen et al., 2013). Fourth, picture choices in the wake of a metaphoric reference come with a cost in terms of response times when compared to those that come after a synonymic one. This finding too is consistent with prior work (e.g. Noveck, 2001). Note that these results follow from the same exact target expressions; it is the context that changes slightly. 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 an in-class experimental tool while also revealing its ability to depict metaphor comprehension development over the school age years.

The older children were associated with performance changes over the course of the task. Notably, analyses of item position showed that the older children's picture choices made in the wake of a metaphoric reference came with increased accuracy and shortened latencies as the experimental session wore on. Interestingly, we also observed changes over trials when older children processed synonymic target references. That is, the older group's correct picture choice time increases in the Synonymic Reference condition as trial numbers do. We surmise that, based on their growing experience with the task and its multiple metaphoric items, these children progressively become aware of dual meanings which prompts them to become more discerning, even among the vignettes that were designed to have a univocal interpretation. Arguably, this explains why the older children's rates of accurate responses in the Synonymic Reference condition

are lower than the younger children's. For the older children, errors in the Synonymic Reference condition reflect the salutary efforts that come from choosing pictures accurately in the Metaphoric Reference condition. Generally speaking, an ability to appreciate alternative meanings is what distinguishes the older children from the younger ones.

Experiment 2

Experiment 2 uses the same task to replicate the developmental effects (while further analyzing it) with the additional aim of determining the extent to which two different abilities –ToM and formal language skills – contribute to children's growing metaphoric performance. That is, we investigate interindividual differences in children's metaphor comprehension in terms of age, like in Experiment 1, but this time we also consider other cognitive abilities, which can be used as predictors in statistical analyses.

As we are interested in interindividual differences in metaphor comprehension and, specifically, in the potential role played by other measurable cognitive abilities, we decided to adopt practices from the field of cognitive assessment. In other words, Experiment 2 was designed as an assessment that *includes* the metaphoric reference task. This led to two changes that serve our purposes. The first is that items in Experiment 2 are now interspersed with other assessment items (which serve as fillers for our purposes). This has the advantage of making the children's task appear as more varied and less repetitive, which should also prevent participants from having expectations about the content of any given vignette. Our expectation is that this should address the unexpected trial order effects that were reported in Experiment 1. The second is that we adopted just one of the lists from Experiment 1 and in a single order. This second change is what makes the presentation of the task's items akin to those used in clinical assessments, which standardize both items and their presented order, (e.g., see the gold-standard IQ assessment WISC-V [Wechsler, 2014]). The advantage of this approach is that differences between two children cannot be attributed to a given list or to order effects. We do not expect this approach to adversely affect the main outcomes reported in Experiment 1. Children's accuracy in making picture choices after a metaphoric reference are still expected to progressively improve with age and their reaction times in making correct metaphoric picture choices are still expected to be more time consuming than synonymic ones.

Method

Transparency and openness

All data and analysis code are available at https://osf.io/myr2c/ (Petit, 2024). This experiment design and its analysis were not pre-registered.

Participants

Two hundred and forty-eight children took part in this experiment. The children's ages ranged from 6;0 to 10;11 years old (mean age = 8 years;4 months, *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 reported by parents, children had no neurodevelopmental disorders (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 associated with children's socio-economic environment (see, e.g., Di Sante & Potvin, 2022). To provide social balance, children were recruited from two different French schools, one a private school in an urban economically privileged area (School A) and another a public school situated in a rural and less economically privileged region (School B). To better characterize each child's home environment, we collected, via parental questionnaires, an estimation of family income through the Family Affluence Scale (FAS, Currie et al., 2008), which is based on 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 and then averaged to provide a socio-economic status composite (SES) varying from 0 to 100.

Table 1

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.

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 local ethics committee (CPP Sud-Est I, France, ID RCB 2019-A01721-56).

Materials

Metaphor task. This Experiment employed one of the lists from Experiment 1 and randomly intermixed its 12 items (6 metaphoric, 6 synonymic) among 22 other items of similar structure and response format. The metaphor task was preceded by 3 training items that familiarized children with the response format, which relied on literal material only and provided the participant with direct feedback. To sustain engagement and attention for this longer procedure, 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. 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 content words (whether they be nouns, verbs, adjectives, or adverbs), based on an oral French

corpus, and frequency of both (4) referent and (5) target nouns in school textbooks (see **Table S2.1** in the 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 with 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 appear 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.

Formal language skills. We assessed formal language skills in the form of grammar reception, 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 the description in an aurally presented sentence. For example, to assess passive structures, children are 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 (correct 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 two training items and 27 test items, and provides a score ranging from 0 to 27.

Theory of mind (ToM) task. ToM was assessed with a tablet-based version of the Picture Sequencing Task (Petit et al., 2024), an experimental task initialy 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 four pictures in a sequence (2 points for correctly placing the first picture, 2 points for correctly placing the last picture, and 1 point for each of the intervening 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., one series of pictures depicts how a speeding truck's vibrations prompt a boulder to roll down a hill), four are based on social scripts (e.g., one sequence depicts meeting a friend for a coffee) and four critical sequences involve false beliefs (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 indices provide average scores ranging from 0 to 6. The PST has been used to study ToM skills in various populations, including high vs. low schizotypal healthy participants (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. (2024).

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 on a variety of tasks including the one on metaphoric reference; 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 at least 2.5 SD below their age group's mean in the formal language 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 metaphoric 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 its 2 conditions. To determine external validity, we computed correlations between accuracy in the critical condition of the metaphoric 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 a lower correlation is expected).

We then fitted mixed effects models, i.e., generalized mixed effect models (for accuracy) and linear mixed effects models (for response times) with the *lme4* package (Bates et al., 2015). In order to address our research concerns, we carried out the following two steps. First, we fitted hierarchical models that include the Reference condition as well as single, quadratic and then cubic terms for age as fixed effects, as well as their interactions with the Reference 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 played by the different cognitive predictors on the DV, we dropped the higher order terms for age in order to prevent computational and interpretational issues. We then added the predictors of interest as fixed effects (language 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 * (Language + 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 visualization. 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 each Reference condition. All continuous predictors were centered on their mean with the *scale()* function. The models' assumptions were checked before reporting their estimates. Single effects were assessed with likelihood ratio tests. All post-hoc contrasts were performed with the *emmeans* package (Lenth, 2022) while adjusting for multiple comparisons.

Results

Eleven 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 sample of 236 children for the analysis. For descriptive purposes, **Table 2** presents the distribution of children across age groups and their results on control tasks.

Table 2

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

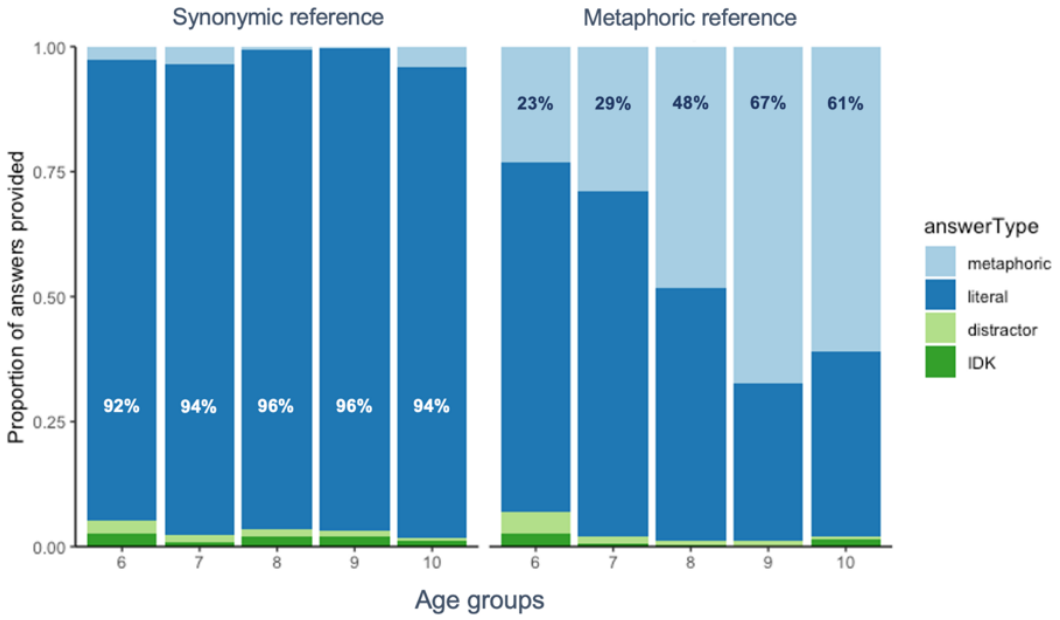
	Groups	6 YO N = 49	7 YO N = 55	8 YO N = 45	9 YO N = 41	10 YO N = 46	*Correlation with age*
Formal language	Mean (*SD*)	13.2 (3.53)	14.5 (2.98)	16.4 (3.71)	17.9 (2.92)	18.3 (3.50)	*r = .51* *p < .001*
TOM	Mean (*SD*)	2.90 (1.01)	3.43 (1.17)	4.13 (1.24)	4.54 (1.10)	4.02 (1.26)	*r = .37* *p < .001*
GSA	Mean (*SD*)	4.44 (1.19)	4.77 (0.82)	4.94 (0.74)	5.47 (0.56)	5.27 (0.80)	*r = .36* *p < .001*
CCC-2 Semantics	Mean (*SD*)	6.6 1.7	6.0 1.6	6.3 1.7	6.2 1.7	6.1 1.4	*r = -.08* *p = .23*
CCC-2 Context	Mean (*SD*)	5.3 2.7	4.9 2.6	5.1 3.4	3.5 2.5	3.9 2.5	*r = -.21* *p = .001*

Task validation

The confirmatory factor analysis clearly supported the bifactorial structure of the metaphor task, one that contrasts synonymic and metaphoric items ($\chi^2(53) = 61$, $p = .21$, *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 Metaphoric Reference condition proved to be excellent without revealing redundancy in the items ($\alpha = .81$, 95% *CI* = [.77, .84]); meanwhile, the internal consistency index of the Synonymic Reference condition appeared lower ($\alpha = .50$, 95% *CI* = [.41, .60]), which could be accounted for by this condition's ceiling effect (see below). Given the ceiling effect, convergent and divergent validity was assessed only for the Metaphoric Reference 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" $r = -.21$, $p = .001$) but not with semantic abilities (CCC-2 "Semantics", $r = .02$, $p = .81$). Finally, the results of Experiment 2 were very consistent with those of Experiment 1: the 7- and 10-year-olds accurately chose the metaphoric representations at rates of 29% and 61%, respectively (in Experiment 1, accuracy rates were 25% and 63%, for the 7- and 10-year-olds, respectively).

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



Accuracy

Accuracy in the Synonymic Reference condition was very high overall (95%), indicating that participants competently chose the intended target. As expected, and as shown in **Figure 2**, rates of correct picture choices in the Metaphoric Reference condition were lower (44%) than in the Synonymic Reference condition. Also, as expected, most errors resulted from the selection of the pictorial representation of the reference taken literally in the metaphoric condition (this accounted for 95% of errors).

As a first step, our analysis revealed that the cubic model outperformed the linear model ($\chi^2(4) = 16.9$, $p = .002$). This model confirmed that participants were much more likely to succeed when comprehending a synonymic reference than a metaphoric one (*OR* = 4.1, *SE* = 0.38, $p < .001$). Post-hoc contrasts revealed no age effects in the Synonymic Reference condition; unlike in Experiment 1, children's performance in the Synonymic Reference condition remained stable across ages. As for the Metaphoric Reference condition, the 6- and 7-years-olds had the lowest rates of accuracy, while being comparable to each other (ß = 0.16, *SE* = 0.3, $p = .93$), the 8-year-olds scored higher than the youngest participants (ß = 1.2, *SE* = 0.2, $p < .001$), and the 9-year-olds outperformed the 8-year-olds (ß = 1.2, *SE* = 0.2, $p < .001$). The 10-year-olds' rates of accuracy were ultimately comparable to those of the 9-year-olds (ß = -0.02 , *SE* = 0.2, $p = .99$).

In a second step, we dropped the cubic and quadratic terms for age and added the additional predictors (see the model's output in **Table 3**). Overall, the model revealed that accuracy in the Metaphoric Reference condition as opposed to the Synonymic Reference condition was predicted by ToM and formal language skills. ToM appeared to have a positive effect on picture choice accuracy in the Metaphoric, as opposed to the Synonymic Reference condition. One also sees that formal language has an influence on metaphoric accuracy compared to the synonymic condition. Each of

these variables and the Reference condition were involved in 3-way interactions with age (age had a moderating effect).

Table 3

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

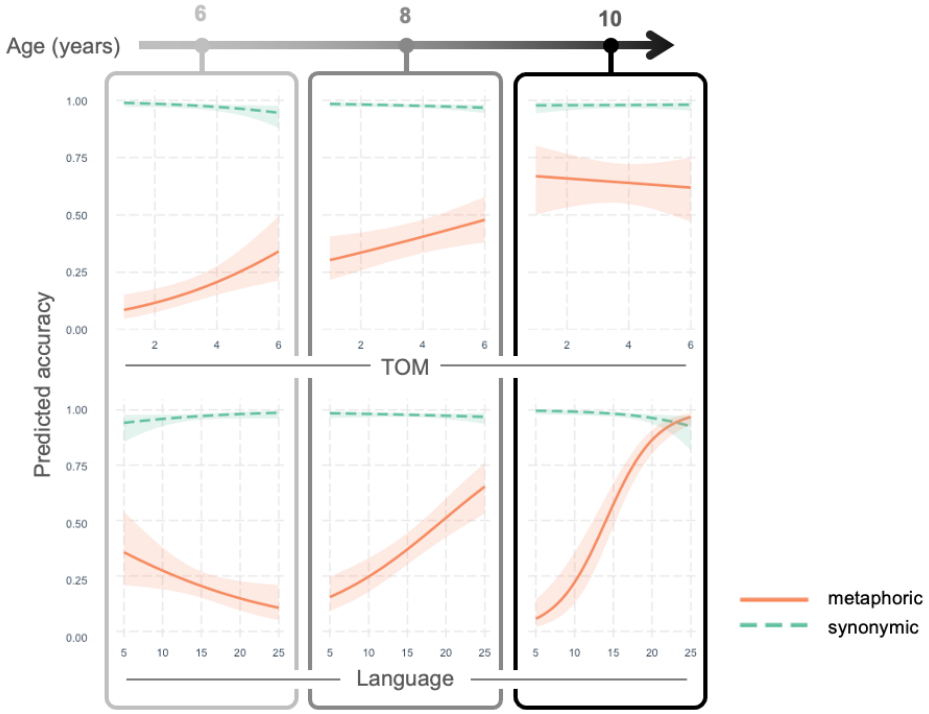
	ACCURACY			LOG(RT)		
PREDICTORS	Odds Ratio	95% *CI*	*p*	Estimates	95% *CI*	*p*
(INTERCEPT)	5.10	3.49 – 7.47	<0.001	8.25	8.03 – 8.47	<0.001
SES	1.20	0.98 – 1.47	0.072	**-0.04**	**-0.07 – -0.00**	**0.023**
AGE	**1.50**	**1.16 – 1.94**	**0.002**	**-0.04**	**-0.06 – -0.01**	**0.008**
CONDITION	**0.01**	**0.01 – 0.03**	**<0.001**	**0.34**	**0.05 – 0.62**	**0.022**
LANGUAGE	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 * LANGUAGE	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 * LANGUAGE	**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) * LANGUAGE	**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**	**0.050**	**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		
OBSERVATIONS	2808			1861		
MARGINAL R^2 / CONDITIONAL R^2	0.493 / 0.648			0.071 / 0.301		

To examine the age moderating effect, the marginal effects of ToM and formal language on accuracy, 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 ToM scores 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 formal language performance predicted higher rates of accuracy in the Metaphoric Reference condition compared to the Synonymic Reference condition at 8, 9 and 10 years of age, though not at 6 and 7 (**Figure 3**, lower half). Interestingly, the only children in the entire sample to produce accuracy rates for the metaphoric reference items that were comparable to those of the synonymic reference items were 10-year-olds who had perfect or near-perfect scores on the formal language test (see **Figure 3**, bottom right).

We note that the ToM index and formal 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 3

Predicted accuracy across conditions, as a function of ToM (x-axis, above) and formal language (x-axis, below), at 3 different ages (6-, 8- and 10-year-olds). Predictions were backtransformed from scaled variables.

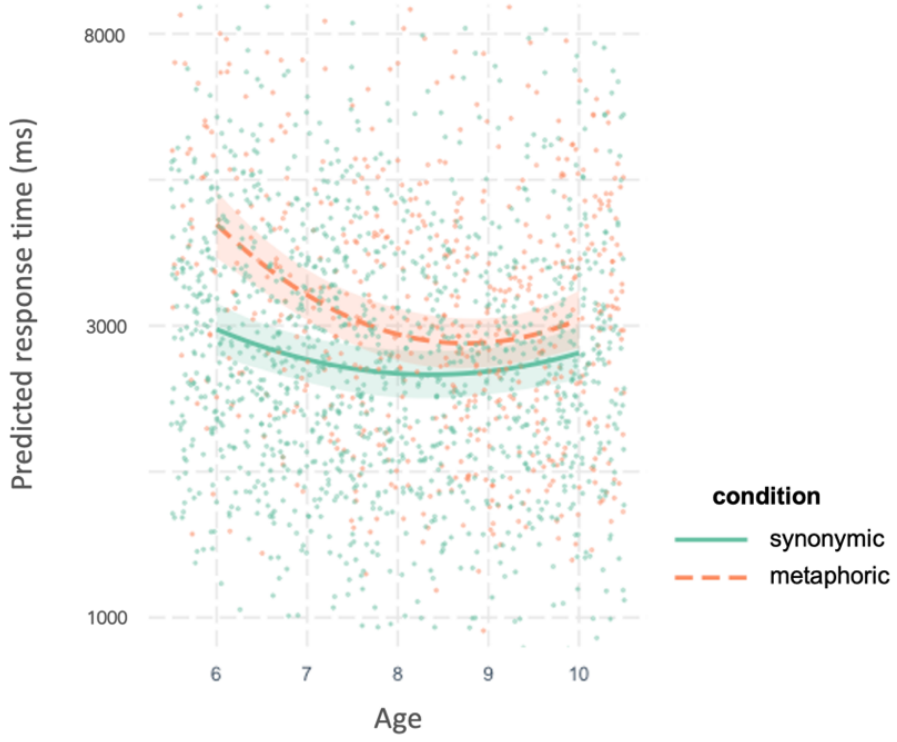


Response times

Before running our analyses on Response times (RT), 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. As a 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 ($\chi^2(2) = 0.7$, $p = .70$). As can be seen in **Figure 4**, the selected quadratic model revealed a main effect of Reference condition, with post-metaphoric choices prompting longer response times compared to post-synonymic ones (ß = 0.13, *SE* = 0.06, $p = .04$). 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.*

Figure 4

Predicted response times (in ms, on a logarithmic scale) as a function of age and condition, and partial residuals (to improve readability, 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).



Reference condition was part of an interaction with age, so one can see that the metaphor-related cost evolved. 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* < .001) and marginally between 8 and 9 years (ß = 0.04, *SE* = 0.02, *p* = .07) but it did not change between 9 and 10 years of age (*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 ToM and formal language were each involved in three-way interactions with age and condition (age has a moderating effect). Higher ToM performance is associated with faster responses in the Metaphoric Reference condition as opposed to the Synonymic Reference Condition; however, this advantage wanes with age. The opposite pattern was observed for formal language.

Post-hoc contrasts (see **Table S2.4** in supplementary materials) reveal that higher scores on ToM predict faster pictorial choice responses after a metaphoric, as opposed to a synonymic, reference at 6 and 7 years of age, but that no such effects were significant at older ages (see **Figure 5**, upper half). Formal language skills are linked to response times of metaphoric, as opposed to synonymic, references at 9 and 10 but not beforehand. Visual inspection of the predictions (see **Figure 5**, lower half) as well as estimates (**Table S2.5** in supplementary materials) indicate that this effect is driven by the language score's negative impact on post-synonymic response times at these ages (leading to longer response times), rather than being due to language's positive impact on response times in the metaphoric condition.

Figure 5

Predicted response times (in ms, log-transformed) across the two reference conditions, as a function of ToM (x-axis, top half) and formal language (x-axis, bottom half), at 3 different age points (6-, 8- and 10-years-old). Note that predictions were backtransformed from scaled variables.

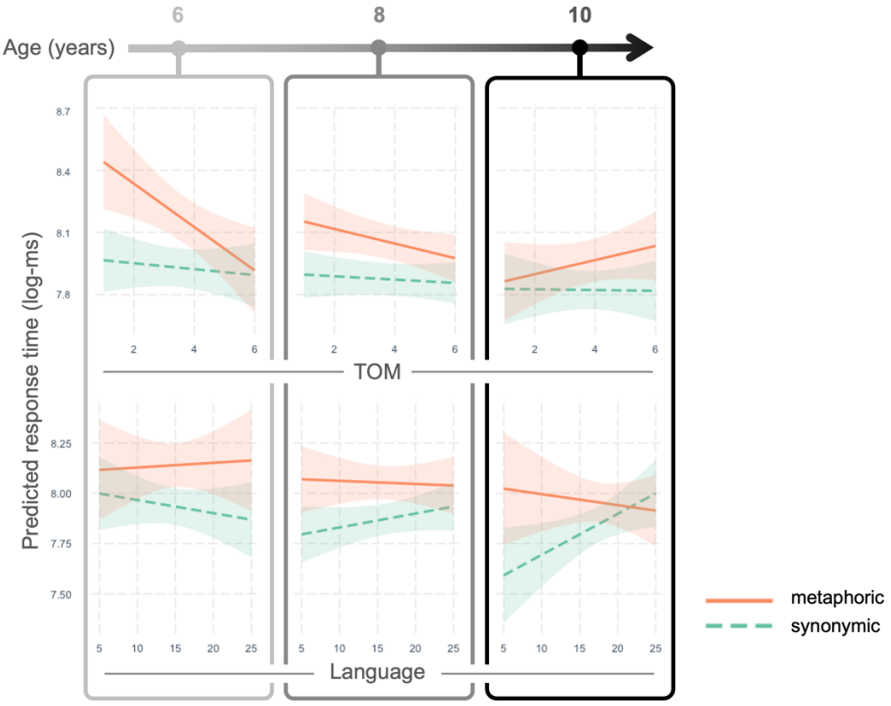
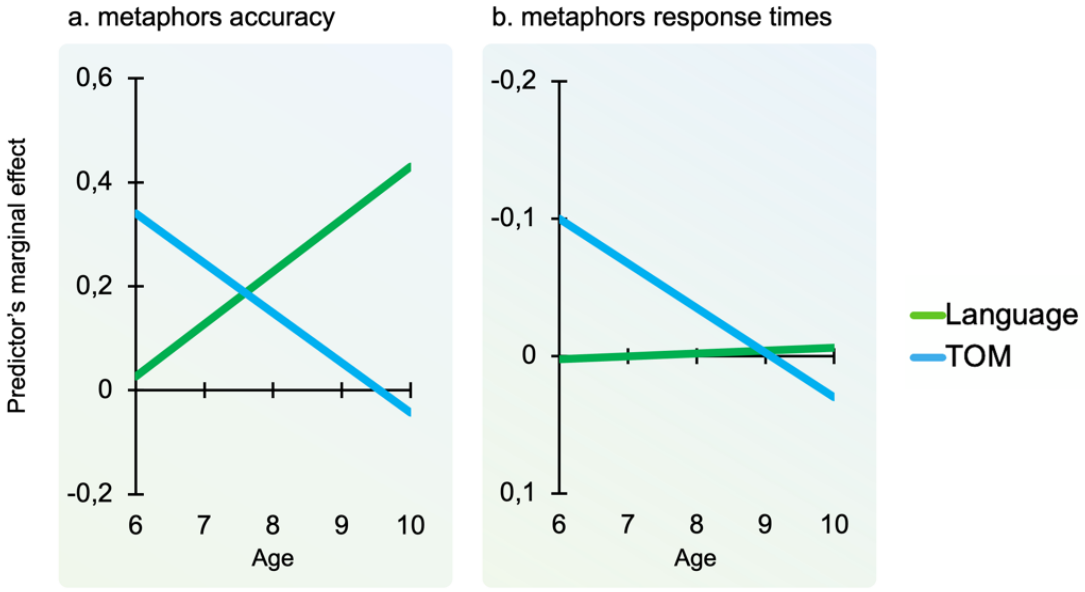


Figure 6 summarizes the marginal effects of ToM and formal language as a function of age on accuracy (left) and response times (right) in the Metaphoric Reference condition. One can see that ToM's influence weakens with age as expressed by both accuracy and response times. Meanwhile, the factor associated with formal language skills facilitates accurate metaphor picture choice (while having little effect on response times) with age.

Figure 6

Predicted partial effects of ToM and formal 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).



Exploratory analysis

To further explore the data and to better 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 – those that reflect literal picture choices in the metaphoric reference condition, which were previously considered as errors. All told, we compared: a) correctly chosen pictures in the Synonymic Reference condition, b) correctly chosen pictures in the Metaphoric Reference condition and; c) incorrectly chosen literal representations of the target reference in the Metaphoric Reference condition. This model showed that the *literal* readings of a metaphoric target reference were made more quickly than correctly chosen pictures post-metaphorically and regardless of age (all *p*s < .001). Moreover, these incorrect choices were made as fast as correctly chosen post-synonymic ones (all *p*s > .05, see **Table S2.6** and **Figure S2.3** in supplementary materials).

Discussion

In Experiment 2, we tested the tablet-based reference task on a second, larger group of children while presenting it as part of an assessment-like scenario. As anticipated, the results with respect to metaphor development were consistent with those found in Experiment 1. Moreover, it provided cleaner findings and finer detail since it included participants across the entire age range, i.e., between 6 and 10 years of age. Most notably, the added fillers in Experiment 2 tempered the one unusual finding in Experiment 1 showing that the younger children's accuracy rates in the Synonymic Reference condition were higher than the older children's. In Experiment 2, there are no drop offs in accuracy across age in this condition.

More germane to our investigation was determining the extent to which the development of ToM, on the one hand, and formal language, on the other, interact with metaphor development. As

far as accuracy goes, the data indicate that the two cognitive abilities interact with development but in two different age windows. Theory of mind appears to have facilitative effects on making accurate metaphoric choices among the younger children. However, these salutary effects appear to attenuate with age. In contrast, the facilitative effects linked with formal language skills appear to not affect performance among the youngest children and to progressively increase with age and maximally among the oldest children. Indeed, the only children who make correct picture-choices after a metaphoric reference – at rates as high as those following a synonymic reference – were the oldest children who scored highest for formal language.

The role of the two cognitive abilities regarding reaction times was similarly revealing. For those children who accurately make metaphoric picture choices, the data reveal that metaphor-related costs evolve with age. That is, the difference between metaphor- and literal-based picture choice response times decreases between 6 and 9 years of age but it does not change between 9 and 10. In contrast, formal language was associated with response-time differences between metaphoric and synonymic references after 9 years of age. Much like in Experiment 1, this appears to be related to increased reflection for post-*synonymic* picture-choices in this age range. That is, the older participants appear to slow down in the Synonymic Reference condition as they become aware of alternative meanings for the target reference (it is not simply the case that there is a speed-up on post-metaphoric picture-choice response times due to a growing sophistication of language skills).

Overall, one can practically see 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 influence appears to wane progressively, ultimately failing to produce significant effects on children by the time they are roughly 8 or 9 years old. At this point, formal language appears to take on an increasingly influential role in children's metaphorical responses.

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, the findings from Experiment 1 provided evidence indicating that metaphor comprehension develops with age while also showing that picture selection based on a metaphoric reference takes reliably longer than its synonymic controls. Importantly for the present paper, the task and findings from Experiment 1 provided us with a proof of concept before it was presented as part of a more comprehensive experiment. That is, in Experiment 2, we repeated the presentation of the metaphoric reference task, but this time with a very large sample of 236 typically developing children, a non-verbal ToM task and a formal language task (as well as items from other tasks that served as fillers here). The findings from Experiment 2 allowed us to further characterize the development of referential metaphors, while considering two cognitive factors that serve as predictors of performance.

Not surprisingly, the data from the metaphor task in Experiment 2 were highly similar to those in Experiment 1. In the remainder of the General Discussion, we consider in greater detail what the task and the results from 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 children's performance on ToM and language skills reveal about metaphor comprehension development.

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

The developmental trend that appeared in both Experiments is consistent with a series of results showing that school age is a key period for metaphor development. As far as control items go, children show little difficulty in 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 task's linguistic material and response format are accessible to children. Meanwhile, picture choice-making in the Metaphorical Reference condition prompted greater equivocality among the children. As children become older their metaphor comprehension performance improves, especially between the ages of 7 and 9 and without particular floor or ceiling effects. At around 8 or 9 years of age, the response times needed to choose metaphoric-based pictures (as opposed to literally-based ones) largely diminishes before stabilizing thereafter. The data also show that incorrect picture choices in the Metaphoric Reference condition typically reflect *literal* interpretations of the target reference.

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 are somewhat lower. This can be explained by three unique features stemming from the paradigm. First, given that we employed oral, instead of written, materials, we were able to test children without concerns about their reading ability. Second, the task includes a multiple-choice response format (instead of, say, a comprehension question requiring a *yes* or *no* response), which yields more variability. This feature had the added advantage of allowing us to provide literal interpretations of metaphoric references as picture choice options (see Van Herwegen et al., 2013) and to positively identify a large proportion of participants' errors as non-random. Third, we based our findings on materials that ultimately presented a single target reference. It was the context that determined whether it was metaphoric or synonymic. Ultimately, the materials were, in our view, more reliable than those used in earlier studies.

This study differed from prior ones in that latencies did not rely on the time needed to *read* a critical sentence that could contain a metaphoric or a literal reference (e.g. Noveck et al., 2001). While these prior studies have the advantage of directly measuring the effort associated with the online processing of the reference, they could arguably be confounded with other reading-related procedures. In our task, we collect response times during a second (picture selection) phase that is downstream from processing the target reference. Interestingly, we still report metaphor-related slowdowns. This shows that a) effects related to metaphor slowdowns are robust and that; b) they are not necessarily related to referring *back* to a previously mentioned element. This indicates that (contra Seigneuric et al., 2016) at least part of successful metaphor comprehension relies, not on reference-making itself but, on what takes place afterward. Note that we are not arguing that metaphor-related slowdowns are merely costly; it is our view that metaphors come with benefits by helping with encoding and memorization (Noveck et al., 2001; Reynolds & Schwartz, 1983).

As far as the current study's qualitative results are concerned, we point out how they are consistent with findings from both classical (Winner et al., 1976) and more recent studies (Deckert et al., 2019; Lecce et al., 2019). These studies generally show that metaphor comprehension development is an ongoing process throughout the school-age years. 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 skills provide scaffolding for metaphoric reference comprehension?

We now turn to our main question, which concerns the potential roles of ToM and formal language skills in metaphor development. Before we begin, however, we make two points regarding our methodology. First, recall that the ToM measure we included employs a minimally verbal procedure (based on Langdon & Coltheart, 1999) that we consider ideal for the purposes of the current study (again, children choose pictures and put them in an order based on a simple instruction). This is unlike the ToM measures used in other studies in this literature, which typically employ assessments based on verbally rich texts or instructions. To make our point, consider the oft-used Strange Stories task (Happé, 1994) as employed by 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. 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 rich vocabulary that distinguishes between words such as *serious, ashamed, alarmed, bewildered* and so on. These sorts of tasks make it potentially difficult to disentangle the respective roles of Theory of Mind and language, which is central to our goal here (see de Villiers, 2007 on the interface between language and TOM). The tablet-based Picture Selection Task used here advances our understanding of the three-way relationship between metaphor comprehension, language and ToM abilities, as called for by Matthews et al. (2018). Second, we use age as a continuous predictor in our statistical analysis, in interaction with the other variables, which allows us to exploit an entire sample's size and variability. This implies that one should view our results as changes that occur during school age, rather than as convergence on exact turning points.

As far as ToM is concerned, one can see its impact on metaphoric reference making starting at 6 years of age, whether one is considering accuracy or response latencies. For example, higher scores on ToM measures are associated with greater accuracy in the Metaphoric Reference condition among the youngest children and this is maintained among children until they are roughly 9 years of age. This finding – showing that the impact of ToM is critical to metaphor comprehension among the youngest children here before it wanes – is in line with findings from Lecce et al.'s (2019) study as well as with Tonini et al.'s (2023). Given the nature of the current study – which relies on referential metaphors, a largely aural presentation, a multiple-choice format, and a ToM measure gathered from a minimally verbal task – one can argue that the prior claims can be generalized and even extended to younger ages (the youngest children in the two prior studies were slightly older than ours). Remarkably, one can readily notice ToM's influence on latencies as well. Higher scores on the ToM task are associated with faster picture choice times among even the youngest children here. This effect, too, falls below significance level by the time children are nine. Overall, the work here confirms and extends the hypothesis that says ToM has a passing influence on metaphor comprehension.

Formal language skills, as evaluated through measures of receptive grammar, also reveal themselves to be statistically predictive of metaphorical reference behavior, even if this association manifests itself with respect to accuracy only. This 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 evolving formal language skills. This pattern resonates with Deckert et al. (2019) who observed (with the Triad task) that verbal intelligence was

associated with superior metaphor performance among 9 and 10-year-olds, but not among 7- or 8-year-olds.

What accounts for these two adjoining developmental trends? Are they linked? We consider two possibilities. The first is that children need to build strong ToM skills in order to understand metaphors (and arguably other related phenomena) in the first place before they can exploit those skills to make refined linguistic judgements. Another possibility is that the reported effects rely on certain linguistic skills that are measurable only among the older children. That is, the items used to measure grammar among the younger children are not the same as those used to distinguish abilities among the older ones. Simple items (for example, those targeting children's understanding of past versus present tense) would reveal distinctions among younger children but not older children, whose refined abilities become apparent through complex structures (such as embedded clauses). Measures reflecting grammatical competence with such complex structures, i.e. those available to older children, are arguably better positioned to be linked with procedures required for metaphor comprehension. These two possibilities are not mutually exclusive.

The apparent developmental shift from ToM-supported metaphor comprehension among younger children to language-boosted comprehension among the older children has implications for both theory development as well as for applied settings. Regarding pragmatic theories in general, our results question a dichotomy recently formalized in the literature between linguistic pragmatics and social pragmatics. According to Andrés-Roqueta & Katsos (2017; 2020), linguistic pragmatics concerns pragmatic cases that depend on structural language and pragmatic norms whereas social pragmatics relies on those plus competence with ToM. In this framework, scalar implicature, a pragmatic phenomenon in which the use of a weak expression justifies the rejection of a more informative one (consider how *some* often prompts *not all*), is considered exemplary of linguistic pragmatics while irony, say, is considered exemplary of social pragmatics.[2] However, our results show that adult-like competence with a single pragmatic phenomenon, metaphor comprehension, can be associated with both ToM *and* language, but at different times of development. While we are sympathetic with the authors' quest to disentangle the umbrella term "pragmatic inference" (for a large review see Noveck, 2018), our developmental results indicate that dividing up the field into phenomena is not necessarily the way to do so. In any case, the current results underline the value of taking a developmental perspective when building and testing models.

The developmental shift we report also opens interesting perspectives for *interventions* aiming to promote metaphor comprehension. The results indicate that either language-based or ToM-based approaches could be used in educational or clinical settings, but that these choices could depend on the age (and undoubtedly other characteristics) of the targeted population. This is important because metaphor comprehension is a useful and ubiquitous communication tool (Bowes & Katz, 2015; Gibbs, 2008; Sopory & Dillard, 2002), which is of importance to children's lives; metaphor comprehension has for instance shown longitudinal and bidirectional associations with peer rejection in childhood

[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, e.g. in certain situations, the linguistic pragmatic analysis of scalar implicature might rely on ToM abilities and vice-versa.

(Del Sette et al., 2021). Moreover, intervention studies could also help specify what is behind the causal nature of the associations we are focused on (Del Sette et al., 2024).

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. We make three points in this vein. The first is that the present work has considered metaphor 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 impact on understanding the latter (Lecce et al., 2019). Second, many prior studies employ vocabulary measures as an index for language skills, while we used grammar. This means one should be cautious when making comparisons among studies that measure linguistic abilities. It is in our collective interest to consider the specific contribution of sub-components of language when investigating metaphor. Third, Experiment 2 relies on a cross-sectional correlational study; the developmental trends we observe should be confirmed by longitudinal studies as well (see Del Sette et al., 2020).

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 metaphoric reference task to the literature, whose outcomes – like its predecessors – show that metaphor comprehension advances with age. Second, the current task, unlike prior ones that largely focus on the target reference itself, finely separates reference assignment from its dependent measure (picture selection) downstream, showing that slowdown effects linked to metaphor comprehension are robust (cf. Noveck et al., 2001). Thirdly, to our knowledge, this is the first study to consider how both ToM and formal language abilities potentially influence metaphor comprehension in a single developmental sample. Specifically, the data show that, while younger children capitalize on ToM, older children's improvement relies on their advancing skills in formal language. Fourthly, the ToM measure employed in the current paper is based on a non-verbal task, making our claims regarding ToM less language-dependent than those drawn from previous studies. Finally, the task is carried out on a tablet that can be one of dozens carried into a classroom. This experimental approach engages young participants and increases the task's ecological validity.

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.

Bowes, A., & Katz, A. (2015). Metaphor creates intimacy and temporarily enhances theory of mind. *Memory & Cognition*, *43*(6), 953–963. https://doi.org/10.3758/s13421-015-0508-4

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. (2023). Metaphor processing: Referring and predicating. *Cognition*, *238*, 105534. https://doi.org/10.1016/j.cognition.2023.105534

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. (2019). 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

Del Sette, P., Bambini, V., Tonini, E., & Lecce, S. (2024). Are theory of mind and metaphor comprehension causally related? A training study in middle childhood. *Language Acquisition*, *0*(0), 1–21. https://doi.org/10.1080/10489223.2024.2345324

Del Sette, P., Ronchi, L., Bambini, V., & Lecce, S. (2021). Longitudinal associations between

metaphor understanding and peer relationships in middle childhood. *Infant and Child Development*, *30*(4). https://doi.org/10.1002/icd.2232

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

Evans, M. A., & Gamble, D. L. (1988). Attribute saliency and metaphor interpretation in school-age children. *Journal of Child Language*, *15*(2), 435–449.

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

Gibbs, R. W. (2008). *The Cambridge handbook of metaphor and thought*. Cambridge University Press.

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

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

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},) [Computer software]. 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

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. (2001). When children are more logical than adults: Experimental investigations of scalar implicature. *Cognition*, *78*(2), 165–188. https://doi.org/10.1016/S0010-0277(00)00114-1

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

Petit, N., Noveck, I., Baltazar, M., & Prado, J. (2024). Assessing Theory of Mind in Children: A Tablet-Based Adaptation of a Classic Picture Sequencing Task. *Child Psychiatry and Human Development*. https://doi.org/10.1007/s10578-023-01648-0

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

R Core Team. (2022). *R: A language and environment for statistical computing* [Computer software]. 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

Rubio Fernandez, P. (2007). Suppression in Metaphor Interpretation: Differences between Meaning Selection and Meaning Construction. *Journal of Semantics*, *24*(4), 345–371. https://doi.org/10.1093/jos/ffm006

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

Sopory, P., & Dillard, J. P. (2002). The Persuasive Effects of Metaphor: A Meta-Analysis. *Human Communication Research*, *28*(3), 382–419. https://doi.org/10.1111/j.1468-2958.2002.tb00813.x

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 Children's Communication Checklist-2 (CCC-2). Translation, adaptation and conceptual equivalence. *Revue*

canadienne d'orthophonie et d'audiologie, *37*(2), 156–168.

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

Wechsler, D. (2014). *Wechsler intelligence scale for children* (5th ed.). Pearson.

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. (2019). 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