Bilingualism: Language and Cognition

cambridge.org/bil

Theory-of-Mind development in educational bilingualism: Identifying the strongest predictors of performance and tracking them over time

Gloria Chamorro[1], Vikki Janke[2] and Inés de la Viña[1]

[1]Universidad Nacional de Educacion a Distancia, Spain and [2]University of Kent, UK

Research Article

Cite this article: Chamorro, G., Janke, V. and de la Viña, I. (2026). Theory-of-Mind development in educational bilingualism: Identifying the strongest predictors of performance and tracking them over time. *Bilingualism: Language and Cognition* 1–14. https://doi.org/10.1017/S1366728926101229

Received: 28 January 2025
Revised: 4 March 2026
Accepted: 4 March 2026

Keywords:
Theory of Mind; educational bilingualism; longitudinal; cognitive; affective; conative

Corresponding author:
Gloria Chamorro;
Email: gchamorro@flog.uned.es

Abstract

This longitudinal study monitored Theory-of-Mind development in monolingually raised but bilingually educated Spanish children (age 5–6) with varied L2-English curricula (13%–83%) to assess whether higher L2-exposure resulted in advantages on seven ToM concepts (emotion, desires, belief, reference, moral-reasoning, lies, sarcasm). Attention (selective, switching, inhibition) and a full suite of individual-difference effects were also monitored. GLMMs linked greater L2-exposure to higher ToM accuracy, and although all three attention measures contributed to ToM scores, the effect of selective attention was the strongest. L1-vocabulary and NVR routinely predicted ToM scores, and girls surpassed boys on sarcasm. We conclude that bilingualism spurs ToM development quickly and is not linked to L2-vocabulary at this stage. In addition, the fact that L2-exposure and individual differences impacted cognitive, affective, and conative ToM differentially supports an approach that analyses these components separately.

Highlights

- The impact of L2 exposure on ToM is visible after one academic year
- The impact of L2 exposure on ToM is independent of L2-vocabulary
- Selective attention, switching, and inhibition contribute variably to ToM
- L1-vocabulary and non-verbal reasoning are integral to ToM development
- Individual differences impact cognitive, affective, and conative ToM differentially

1. Introduction

The nature of the relationship between bilingualism and children's social, cognitive, and linguistic development continues to be debated. Whereas early literature (Ausubel et al., 1980; Hakuta, 1986) warned that exposure to bilingualism might impede these skills, there is now literature supporting bilingualism's positive effects on cognition (Bialystok & Martin, 2004; Cape et al., 2018; Costa et al., 2008; Genesee, 2004; Hernández et al., 2013) and literature that does not (Antón et al., 2019; Branzi et al., 2016; Hernández et al., 2013; Paap & Greenberg, 2013; Paap et al., 2015). Studies have become more precise about the cognitive skills that bilingualism might enhance (Costa et al., 2009; Bialystok & Craik, 2010) and have discussed the intensity and type of bilingual environment in which positive effects are more likely to materialize (Bialystok & Magumuder, 1998; Carlson & Meltzoff, 2008; Hermanto et al., 2012; Kalashnikova & Mattock, 2014). Focus has also turned to the potential contributions of myriad variables present in this group, such as socioeconomic status (SES), caregivers' educational background, immigrant status, first-language (L1) proficiency, and working memory (WM) (Chamorro & Janke, 2020; Engel de Abreu et al., 2012; Hughes et al., 2005; Nguyen & Astington, 2014; Paap et al., 2015). Still more recently, researchers have examined whether early-observed advantages continue longitudinally (Chamorro & Janke, 2023; Chamorro et al., 2025; Dick et al., 2019; Nichols et al., 2020; Rubio-Fernández & Glucksberg, 2012).

These questions extend to research on bilingual children's social cognition – the main focus of our study – in particular, their developing Theory-of-Mind (ToM) (Premack & Woodruff, 1978). ToM refers to the ability to understand others' beliefs, desires, and thoughts and to recognize that these will influence their behavior (Wellman, 2018). As ToM tasks exploit conceptual (constructing several representations) and executive-function (EF) skills (maintaining and toggling between multiple representations and discarding one in favor of another), it is unsurprising that bilingualism might boost aspects of ToM too (Bialystok & Senman, 2004; Farhadian et al., 2010; Goetz, 2003; Kovacs, 2009; Schroeder, 2018). Some studies have reported overlaps in neurological activation in participants undertaking ToM and EF tasks (van der Meer et al., 2011), warranting further comparisons between patterns of performance on these different batteries.

There is, for example, research that points to EF having a positive effect on ToM in monolinguals (Devine & Hughes, 2014; Doenyas et al., 2018; Hughes & Devine, 2019), and work that has shown EF development to predict ToM development but not vice versa (Carlson et al., 2002), which might suggest that EF skills underpin those utilized in ToM (Markovitch et al., 2015). However, the overlap is not necessarily neat and may depend on the ToM task administered (Schlaffke et al., 2015; Sebastian et al., 2012).

With respect to bilinguals, Huang et al. (2023) linked non-verbal WM and cognitive flexibility – but not inhibitory control – to ToM scores in Spanish-English children from low-income families. This and other work have urged researchers to extend the breadth of ToM tasks administered beyond the prototypical cognitive false-belief (FB) tasks (Wellman, 2018) and to monitor the influence of other variables, such as L1-proficiency, on ToM (Białecka et al., 2024), a factor strongly associated with ToM in monolinguals (de Villiers & de Villiers, 2014; Milligan et al., 2007).

In this longitudinal study, our primary question is whether children with higher L2-exposure achieve higher ToM scores. We respond to concerns over the heterogeneity in bilingual groups and the call to utilize different ToM tasks by focusing on bilingually educated children, whose access to bilingualism is restricted to school, and by administering an extensive battery, which assesses understanding of desires, emotions, beliefs, reference, moral-reasoning, lies, and sarcasm (Sotomayor-Enríquez et al., 2023). These concepts comprise either chiefly cognitive, affective, or conative components (see below), enabling us to check if performance on these is differentially affected. However, limiting the present participants to those from monolingual homes does more than introduce predictability in terms of L2 exposure. It means that their access to the broader social interactions familiar to multilingually and multiculturally raised children is reduced. Since the cultural diversity that bilingual children experience can affect their sensitivity to others' emotions (Bukhlaenkova et al., 2022; Cheung et al., 2010), minimizing this diversity diminishes another potentially influential factor. In addition, the intensity and duration of second-language (L2) exposure is important, so by including L2-exposure as a continuous variable, we monitor these influences on ToM too. As a secondary question, we explore how three attention measures (selective, switching, inhibition) relate to ToM, and finally, we track the influence of a suite of individual differences (non-verbal reasoning [NVR], WM, L1-vocabulary, L2-vocabulary, age, gender, family education, other language [s] at home, onset of L2-exposure, L2-exposure outside school, and exposure to further languages beyond school). Our introduction starts with literature that has focused on bilingualism in a broad range of bilingual contexts and its relation to ToM, before turning to investigations that narrow down to educational bilingualism and ToM. Where space permits, we include EF literature that relates directly to this paper's chief ToM focus. The current paper complements Chamorro et al. (2025), who examined educational bilingualism and attention, and found that higher L2-exposure and higher L2-vocabulary consistently predicted higher selective attention, switching, and inhibition scores. It will be particularly interesting to see, therefore, whether the variables that predicted attention scores on this population also predict ToM scores.

1.1. Bilingualism and ToM

This section presents studies of participants with different language histories, highlighting how their degree of dominance is reported. Where information is available, we include which aspects of ToM were measured to see if distinctions emerge with respect to the tasks' cognitive, affective, or conative components (Shamay-Tsoory & Aharon-Peretz, 2007). When reported, EF measures that were compared with ToM will also be signaled.

The majority of studies on bilinguals' ToM development has focused on FB (Yu et al., 2021), which prototypically involves a child being presented with a vignette where they see that, unbeknown to a protagonist, the location/content of an item changes. The task tests whether children grasp that the protagonist will continue to falsely believe that the original location/content is correct (Astington & Gopnik, 1991; Wimmer & Perner, 1983). By 4 years, most children pass simple versions (Wellman et al., 2001). FB tasks are categorized as cognitive as they require children to hold two mental representations (one corresponding with reality and the other not) and to choose the one that conflicts with reality over the one that does not (Carlson et al., 2002). In contrast, affective-ToM tasks test whether children identify emotions, distinguish overt emotional expressions from actual emotional states, and feel compassion for others (Dennis et al., 2013; Feng et al., 2023). Like their strictly cognitive counterparts, affective-ToM tasks incorporate mentalizing, but the additional inferences drawn concern emotions, not beliefs or knowledge. In light of Shamay-Tsoory et al. (2007), who reported a dissociation between cognitive and affective-ToM in adults with contrasting brain traumas, it is possible that these skills are differentially influenced.

FB tasks become more complex by introducing a second protagonist into the vignette and asking 'what does X believe that Y believes about a circumstance'. Performance on these second-order tasks develops later, from around six (Liddle & Nettle, 2006); possible reasons, beyond manipulation of more representations, are that they are more complex linguistically since their explanation involves several layers of sentence embeddings. This increased linguistic complexity makes their results harder to pin to purely cognitive factors. Other cognitive-ToM tasks include physical appearance-reality tasks (Carlson et al., 2002; Flavell et al., 1983), where the challenge is to distinguish between what something looks like and what it actually is, and visual perspective-taking tasks, which test the ability to recognize that an item will look different from alternative viewpoints.

Bialystok and Senman (2004) examined the relation between bilingualism and performance on an appearance-reality task, expecting 4- and 5-year-old bilinguals to outperform monolinguals on the basis of their higher inhibitory-control scores. The bilinguals had families with recent immigrant status, a range of language backgrounds, spoke L2-English in their schools/neighborhoods, and L1 at home. The degree of dominance or age at which they started to learn English is not stated. Shown 'real' objects (a sponge looking like a rock) and 'representational' ones (a whale functioning as a pen), children were asked what they were. On answering, the objects' true characteristics were revealed. The questions included (a) what the children first thought the item was, (b) what someone who had not been shown its true characteristics would think it was, and (c) what it really was. With language controlled, bilinguals outperformed monolinguals on (c). In addition, for all children, composite inhibitory-control scores correlated with performance on (c), which made sense conceptually since to succeed on both tasks they had to suppress a salient yet irrelevant representation in favor of a correct one. This finding maps with literature on monolinguals' ToM, which has linked inhibitory control (alongside other EF skills) to successful execution of ToM tasks (Apperly, 2011).

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

Other work has pointed to the role of WM for ToM in bilinguals. Nguyen and Astington (2014) examined the relation between WM (Backward Word Span), conflict inhibition (Day/Night, Stroop), and FB (unexpected contents/location). It compared 3- to 5-year-old English and French monolinguals with English-French bilinguals. The latter had been exposed to English and French before 8 months and were exposed to both languages minimally 30% of the time. With bilinguals' lower vocabulary scores controlled, they outperformed monolinguals on FB and WM but not conflict inhibition. There was also a correlation between FB and WM, which was attributed to WM exploiting similar skills as FB, such as maintenance and manipulation of items.

Bilingualism also predicted performance on a spatial-perspective task in Greenberg et al. (2013), where the verbal element to the task had been greatly reduced in an effort to pin the bilingual advantage more confidently to EF skills. Eight-year-old bilinguals (representing 15 languages), 89% of whom spoke their non-English language at home and 62% of whom spoke English and their other language daily, were compared with monolinguals. The groups did not differ on English vocabulary, NVR, or SES, and it was bilingualism that predicted performance on this purely cognitive task, arguably due to their better EF skills.

Other studies have found ToM performance in bilinguals to be related to language abilities rather than cognitive ones. Diaz and Farrar (2018) compared Spanish-English bilingual and monolingual 3- to 5-year-olds on FB and appearance-reality. The language histories of the bilingual children were diverse: 61% were exposed to both languages from birth, spoke regularly in both, and those whose access to their least dominant language had come later, had been exposed to it for at least a year. With language controlled, a bilingual advantage was found again for (composite) ToM. However, whereas general language ability predicted composite ToM scores for both groups, interference suppression and WM mediated monolinguals' scores. As the FB and appearance-reality tasks were not analyzed separately, one cannot tell if different factors predicted performance on them. In a longitudinal version of this study, Diaz and Farrar (2017) compared 3- to 5-year-old Spanish-English bilinguals with monolinguals. Again, the bilinguals had varied language histories, with 78% experiencing both languages from birth and the remaining being exposed to their non-dominant language for at least a year. With vocabulary controlled, bilinguals outperformed monolinguals on composite ToM (and EF) scores at first testing. Expressive vocabulary and metalinguistic awareness predicted bilinguals' ToM scores, whereas interference suppression and metalinguistic awareness predicted monolinguals'. One year later, the group differences on FB or EF had gone, and performance predictors had changed: metalinguistic scores (where the key trial was a synonym task) predicted bilinguals' ToM but language (receptive language and complementation) and cognitive-flexibility (Dimensional Change Card Sort, DCCS) scores from the first testing phase predicted monolinguals' ToM. Again, use of composite ToM scores prevents any disentangling between FB and appearance-reality with respect to these predictors.

Białecka et al. (2024) looked further at the role of language proficiency on ToM but included an affective component in their tasks. They compared Polish-English 4- to 6-year-old immigrant bilinguals (living in England but also spending time in Poland), with Polish monolinguals. The bilinguals' age at onset of L2-exposure varied substantially and the Polish-English balance at home is not stated. Of the ToM tasks tested, most were cognitive, but some considered feelings, thereby incorporating affect into the composite score. Children gained points for accuracy and justification of answers, and whereas no differences emerged for accuracy, bilinguals did better with justifications. For all groups, L1-proficiency predicted accuracy but for monolinguals, NVR did, too. However, the strongest predictors for bilinguals' justification scores – the harder task cognitively and linguistically – were L1- and L2-vocabulary.

Han and Lee (2013) compared cognitive and affective perspective-taking abilities in 4- to 5-year-old bilinguals (Korean-English) and monolinguals (Korean). For the cognitive part, children retold a story from another's perspective, and for the affective, they decided on a character's emotions. The bilinguals had comparable Korean and English vocabulary scores. All children lived in South Korea with Korean parents, but the bilinguals were largely born outside of South Korea, attended international kindergartens, and had varied experiences living abroad. Their finding, that bilinguals outperformed monolinguals on the affective but not the cognitive task, is difficult to interpret because scores on the affective task were very high overall, whereas 4-year-olds scored less than 50% correct on the cognitive one, indicating substantial difficulty with it. NVR, WM, or other EF skills were not measured, so the study cannot speak to whether these factors contributed to overall performance or whether cognitive and affective components were differentially affected. Note, for example, that in Cassetta et al. (2018), monolinguals' performance on inhibitory control and switching predicted (second-order) cognitive-ToM but not affective-ToM. However, Han and Lee's results might suggest that the generalization that cognitive-ToM precedes affective-ToM, as per Shamay-Tsoory et al. (2007), is too strong and dependent on task difficulty. Success with cognitive second-order FB tasks does seem to precede success with social-faux-pas tasks (Baron-Cohen et al., 1999), where the latter includes complex cognitive and affective components. Keeping track of the type of ToM task administered is important as the ability to infer thoughts differs from the ability to infer feelings (Healey & Grossman, 2018), and in other monolingual adult groups, these abilities are associated with different EF skills. Healey and Grossman, for example, found switching was more relevant to cognitive than affective perspective-taking. Viewed in relation to the literature demonstrating how development of inhibitory control (Grote et al., 2021; Verhagen et al., 2017) and switching (Planckaert et al., 2023), in particular, favor the bilingual child, this is another reason that monolinguals' and bilinguals' performance on cognitive and affective-ToM might diverge.

To summarize, we have reviewed the performance of bilingual children with quite disparate language histories on a range of ToM tasks and found that different linguistic and EF predictors emerge not just between monolinguals and bilinguals but between different bilingual types. Given the heterogeneity in bilinguals across studies, a question that arises is whether results could become clearer by forging greater homogeneity in this group. Focusing on educational bilingualism – where L2-access starts at pre-school/primary-education (PE) and is largely restricted to this environment – is one way of extracting a more homogeneous subset, and there are only a few studies that have examined ToM in relation to this group.

1.2. Educational bilingualism and ToM

Goetz (2003) provided an example of educational bilingualism benefitting ToM development. It compared two monolingual preschooler groups (Mandarin, English) with Mandarin-English bilinguals whose access to English began in daycare. Children participated twice (T1, T2) on three cognitive tasks: appearance-reality, perspective-taking, and FB (unexpected contents/transfer). There was a bilingual advantage for overall ToM at T1 but

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

not T2, and for all children, L1-vocabulary correlated positively with ToM at T1 but not T2. A key difference between groups with respect to T1 and 2 was that monolinguals repeated the tasks in the same language, whereas bilinguals completed them in English and then in Mandarin. As Goetz suggests, this might account for why monolinguals' performance improved, whereas bilinguals' remained constant. However, it is also possible that an initial boost provided by bilingualism faded quickly. Goetz's bilingual group does not map sufficiently tightly to ours because although Mandarin was the primary home language, the children had some access to English via family friends and activities such as TV. Buac and Kaushanskaya (2020) accords more closely in this respect. Focusing on 7-year-old children, it compared monolinguals with two bilingual groups with different language histories on an English second-order FB task. One bilingual group had access to two languages before age three, but the other group had only L1-English at home and were introduced to Spanish at school from six, where the language balance was typically 70% English to 30% Spanish. The simultaneous bilinguals struggled most with FB and the factors predicting the groups' performance differed. For simultaneous bilinguals, whose English was significantly lowest, expressive English language skills predicted FB. For monolinguals, it was WM, yet for the immersion group, interference inhibition and switching scores were the relevant predictors. One possible interpretation is that the importance of language for FB is central when that language is (relatively) impoverished, but as it improves, aspects of EF take over as key. Comparison between the simultaneous bilingual and immersion groups on FB in Spanish, where language abilities would not have been at stake, might have clarified this.

Cheung et al. (2010) compared FB in 3- to 4-year-old children with the same home language (Cantonese) but with different access to L2-English. Both groups had monolingual families, but one attended an immersion kindergarten, with all activities in English. The other attended a Cantonese-speaking kindergarten with 5 hours of English per week. After 1 year, the immersion children fared better on FB, suggesting that there might be a minimum threshold for L2-exposure to impact ToM, 5 hours per week being insufficient. However, the authors linked the difference to sociolinguistic awareness: the immersion children switched languages completely between kindergarten and home, and it was this difference that propelled sociolinguistic awareness, and in turn, ToM. Indeed, it was sociolinguistic awareness scores that predicted FB performance. Agostini et al.'s (2025) results were more promising with respect to the linguistic threshold for positive ToM effects. They compared 4- to 5-year-old monolingual, bilingual, and L2-learner groups on perspective-taking in a communicative setting on starting school and 6 months later. At first, children's performance did not differ, but 6 months on, the bilingual and L2-groups outperformed monolinguals. However, neither interference inhibition, cognitive-flexibility, nor L1-English vocabulary predicted performance on this cognitive task.

These few studies examining ToM in relation to educational bilingualism demonstrate that further inroads could be made by studying a greater number of children in this less-studied and more homogeneous group and by including school L2-exposure as a continuous variable. A longitudinal study on this group would reveal how quickly any advantages that do materialize occur and how they pattern over time. Finally, documenting children's success on individual ToM concepts rather than on composite scores would enable comparisons between performance on ToM components whose driving force are cognitive, affective, or conative (see Section 2.2). Our study is situated within this context.

1.3. The present study

Our main aim is to compare the ToM scores of bilingually educated Spanish children with monolingual Spanish children longitudinally. We employ Sotomayor-Enríquez et al.'s (2023) task, which builds on the developmental progression of sociocognitive abilities proposed in Wellman and Liu (2004) and includes cognitive, affective, and conative components. A subsidiary aim is to monitor whether attention skills further modulate ToM success. Lastly, our inclusion of a range of individual-difference factors will track their contribution to ToM scores. Our questions and hypotheses are as follows:

RQ1. How does proportion of English at school relate to performance on individual ToM components (desires, emotion, belief, moral-reasoning, reference, lies, sarcasm) longitudinally?

H1. Greater L2-exposure at school will pattern with higher ToM scores at the year's start, and this association will have strengthened at the year's end; however, cognitive, affective, and conative ToM components may be differentially affected.

RQ2. To what extent does performance on attention (selective, switching, response inhibition), beyond that of L2-exposure, contribute to ToM scores longitudinally?

H2. Higher attention scores will pattern with higher cognitive ToM scores, but affective and conative ToM components may be differentially affected.

RQ3. Do individual differences (NVR, WM, L1-vocabulary, L2-vocabulary, age, gender, family educational level, other language(s) spoken at home, onset of L2-exposure, L2-exposure beyond school, exposure to further languages beyond school) influence children's ToM scores?

H3. Higher NVR, L1-vocabulary, and L2-vocabulary scores, in particular, will map to higher ToM scores.

2. Method

2.1. Participants

A total of 231 Spanish children from 10 Madrid schools were recruited.[1] Teachers and parents reported they had no social, cognitive, or linguistic conditions. There were two testing phases: beginning of PE (T0), age 5–6; and end of Year 1 (T1),[2] age 6–7. Children were grouped as monolingual, bilingual, or immersion according to English exposure at school. The 'monolinguals', from four schools (three state, one semi-private), attended non-bilingual schools whose curricula were delivered in Spanish aside from 3 hours of English per week, which amounted to 13.3% of their curricula. The bilinguals, from five schools (three state, two semi-private), attended bilingual schools where 32–41.1% of the curricula were in English. The immersion group, from one British (private) school, had an 82.86% English curriculum, with 4.5 hours of Spanish per week. Table 1 shows each group's number, gender, and age at T1.

[1]School-level recruitment constraints limited our sample size, and no a priori power analysis was conducted (noted in line with transparency recommendations; Lakens, 2022).

[2]At T1, numbers were 229 due to two school leavers.

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

Table 1. English exposure, number, gender distribution, and mean age (SD) by group at T1

	Monolingual	Bilingual	Immersion
% English exposure	13.33	32–41.11	82.86
Number (Gender)	66 (33 girls)	138 (77 girls)	25 (16 girls)
Mean age (SD)	6.68 (0.39)	6.63 (0.40)	7.03 (0.27)

2.2. Materials

2.2.1. ToM measures

For ToM, we used the Theory-of-Mind-Booklet Task (Sotomayor-Enríquez et al., 2023), developed for longitudinal studies with children aged 3–12, which includes two booklets.[3] The task exists in English and German, so we translated it into (Peninsular) Spanish, piloting it first. It comprises stories and pictures (see Figure 1) that describe a protagonist's mental state and appearance, as well as physical events, objects, and states, and that capture stable individual differences and developmental changes. It includes binary forced-choice and free-response questions covering a range of concepts. At T0, we administered Booklet 1, with two stories: one about children finding schoolbooks and another about playing in a park. It contains 42 items assessing five concepts: Desires, Emotion, Belief, Reference, and Moral-Reasoning. At T1, children completed Booklet 2, featuring a story about looking for snacks. By using two booklets designed for longitudinal studies, we tracked changes while avoiding practice effects from using the same stories. Booklet 2 includes the ToM concepts assessed in Booklet 1 but adds Lies and Sarcasm, raising the item total to 50.[4] For each item, children scored 1 if correct and 0 if incorrect. Table 2 illustrates the concepts included and their cognitive/affective/conative categorizations. Cognitive mentalizing refers to ToM tasks requiring cognitive perspective-taking, distinguishing physical appearance from reality and assessment of others' cognitive beliefs. Affective mentalizing includes recognizing emotions behind facial expressions and, at a more advanced stage, understanding deceptive emotional expressions of states. Lastly, conative mentalizing requires an understanding of how and why someone can influence the thoughts/feelings of another, such as by empathizing or using sarcasm/irony (Dennis et al., 2013).

2.2.2. Attention measures

Children completed the Test-of-Everyday-Attention-for-Children (Manly et al., 1999), which is a standardized and normed clinical battery of attention tests. They undertook four tasks (six measures) assessing three attention types: selective, switching, and response inhibition.

Figure 1. False-belief example from Booklet 2.

Table 2. ToM concepts included in each booklet, their score range, and categorization

	Booklet 1	Booklet 2	Categorization
Desires	0–4	0–2	Cognitive
Emotion	0–6	0–8	Affective
Belief	0–13	0–14	Cognitive
Reference	0–6	0–8	Cognitive
Moral-reasoning	0–13	0–11	Cognitive
Lies	*Not included*	0–5	Cognitive
Sarcasm	*Not included*	0–2	Conative
Total items	42	50	

- SkySearch (selective attention: timing): Children locate as many pairs of identical spaceships (targets) as quickly as possible among pairs of different spaceships (distractors). The score reflects timing but takes accuracy into consideration too.
- CreatureCounting (switching: accuracy and timing): Children see rows of 'creatures' with arrows pointing up or down inserted between some of them. They count the creatures aloud, switching the way they count based on the arrows' directions. There are seven trials with two measures: accuracy (number of trials in which creatures are counted correctly) and timing (mean time children take in accurate trials; timing is taken if they get at least three trials correct).
- Walk/Don't Walk (response inhibition: accuracy): A series of tones play, and for each, children mark a step along a path except for the last tone, which ends differently, signaling them to stop unpredictably. The task has 20 trials; the score represents the total number of correct trials.
- OppositeWorlds (switching: timing – congruent and incongruent): Children see paths of numbers, '1'and '2'. In the congruent condition, they read the numbers as they are; in the incongruent condition, they read '1' as '2' and '2' as '1'. There are four trials with two in each condition (i.e., two timing measures in each condition).

2.2.3. Individual-difference measures

Standardized measures included NVR, assessed by Raven's Coloured Progressive Matrices (Raven et al., 1998); WM, assessed by the Forward Digit Span task from the Wechsler Intelligence Scales for Children-Revised (WISC-R; Wechsler, 1974); and L1- and L2-receptive vocabulary, assessed by the Test de Vocabulario en Imágenes Peabody (PPVT-III; Dunn et al., 2006) and the British Picture Vocabulary Scales (BPVS3; Dunn & Dunn, 2009), respectively.[5] Children's age and gender were recorded, and parents completed a questionnaire at T0 on their educational background, immigrant status, home language(s), children's onset of L2-exposure, and amount/type of exposure to foreign languages outside school, which were collected at school.

2.3. Procedure

With school and parental consent in place, children undertook all tasks at T0 and T1. Sessions occurred individually in a quiet room,

[3]This is available at https://osf.io/g5zpv/.

[4]Booklet 2 includes 11 control items on which all groups reached ceiling.

[5]At T1, a Backward Digit Span task was also included.

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

during school, over three 30-minute sessions on different days. During the first, children completed Raven's then Digit Span and then BPVS. In the second, they undertook the ToM task and then PPVT, and in the third, the attention test. Tasks were explained and conducted in Spanish, aside from BPVS which was explained in Spanish and conducted in English.

2.4. Analyses

Analyses were conducted in R (version 4.2.1; R Core Team, 2023) with Generalized Linear Mixed Models using the glmmTMB package. Given that our study's main focus was to identify how development progressed across ToM concepts, the analyses were conducted at the concept level. The dependent variable was Concept-Level[6] score (coded 0/1), which was calculated for the concepts not at ceiling, with 'concept' included as a fixed effect. Concepts exceeding the conservative ceiling criterion (≥ 0.95 accuracy) were excluded from inferential modeling to ensure variability.

Models were fitted at T0 and longitudinally across T0 and T1 for all ToM concepts included at T0 and T1. T1-only models were run for the two concepts introduced at T1 (Sarcasm, Lies). Additional models were fitted to examine the contributions of attention measures (selective, switching, response inhibition) to ToM.

The main predictor of interest was percentage of English-at-school,[7] with time (T0, T1) and their interaction included to explore longitudinal changes. Attention measures were added as predictors in a subset of extended models to evaluate their role beyond that of L2-exposure in explaining ToM outcomes. For categorical predictors (i.e., ToM concept), deviation coding was used. Models also included covariates (L1-vocabulary, L2-vocabulary, NVR, WM, age, gender, family education, further language(s) spoken at home, age of first exposure to English, weekly exposure to English, and other languages outside school).[8] A systematic model selection process was employed, beginning with a full model containing all fixed effects and applying backward elimination based on likelihood ratio tests to simplify models (Plonsky & Ghanbar, 2018). The main predictors of interest (English-at-school, time, their interaction) were retained in final models. For each model, the maximal random effects structure that converged was employed, with random intercepts for subjects to account for repeated measures (Barr et al., 2013). Model assumptions, including uniformity, dispersion, and outliers, were checked through residuals using the DHARMA package (version 0.4.7; Hartig, 2022). Model outputs are in the Supplementary Materials (Tables S5–S11).

[6]These models used a binomial distribution with deviation coding and included random intercepts (and slopes where possible). These assessed whether the effect of English-at-school varied across specific ToM components via the 'concept × English-at-school' interaction.

[7]We had two L2-exposure measures: percentage of 'English-at-school', used in the GLMM analyses to examine fine-grained variations in exposure, and 'Type of School', (monolingual/bilingual/immersion) used only for descriptive overviews of the educational settings.

[8]Preliminary collinearity analysis revealed a moderate negative correlation between immigrant status and educational level ($r = -0.40$, $p < .001$). A relationship was also found between school category (state/semi-private/private) and school L2-exposure ($F(2,459) = 774.4$, $p < .001$). To avoid multicollinearity, immigrant status and school category were consequently excluded as was 'Type of school' (monolingual/bilingual/immersion), given its conceptual overlap with school L2-exposure and its relationship with this variable ($F(2,459) = 774.4$, $p < .001$). See Supplementary Materials (Table S12) for the full correlation matrix.

Table 3. Age, NVR, WM, and L1- and L2-vocabulary means (SDs) by group at T0

	Dutch				English			
	n	*M*	*SD*	Range	*n*	*M*	*SD*	Range
PPVT (raw scores)	63	78	22	32–100	51	79	33	22–152
CLT	60	20	8	4–35	44	18	8	4–35

3. Results

Preliminary analyses of individual- and demographic-difference variables using 'Type of School' as a grouping variable showed groups were matched on NVR, WM, age, and L1-vocabulary. However, there were differences in L2-vocabulary ($\beta = 0.025$, $SE = 0.002$, $z = 9.18$, $p < .001$). Children in immersion and bilingual schools had significantly higher vocabulary scores than those in monolingual schools (see Table 3).

Initial ToM analyses showed children from all three groups scored at or above the 95% ceiling threshold on Desires. Consequently, Desires was excluded from the analyses.[9] Figure 2 shows the percentage accuracy for each concept by group. Descriptively, it suggests a consistent ordering of difficulty across groups, with children performing best on Desires, followed by Emotion, Belief, and finally Reference and Moral-Reasoning.

In the following subsections, we begin with results at T0 to gauge preliminary group differences. This is followed by an analysis of the changes in performance between T0 and T1, once children had completed their first year of PE. We end with the results on attention.

3.1. T0

In our analyses, we modeled Concept-Level-ToM for concepts not at ceiling, using deviation coding for the 'concept' factor. This enabled comparisons of each concept's accuracy relative to the overall mean across all the concepts included. The model resulted in a significant main effect of English-at-school ($\beta = 0.014$, $SE = 0.004$, $z = 3.55$, $p < .001$), with greater L2-exposure associated with increased odds of correct responses. Among ToM concepts, Emotion ($\beta = 1.569$, $SE = 0.170$, $z = 9.22$, $p < .001$) showed significantly higher accuracy than the grand mean, while Moral-Reasoning ($\beta = -0.563$, $SE = 0.188$, $z = -2.99$, $p = .003$) was significantly below it. Belief did not differ significantly from the grand mean ($\beta = -0.111$, $SE = 0.168$, $z = -0.66$, $p = .508$). Because deviation coding was used, Reference cannot be directly estimated in the model. However, its value can be derived as the negative sum of the other concept estimates (Reference = −[Belief+Emotion +Moral-Reasoning]). Based on these estimates ($\beta = -0.111$, 1.569, and − 0.563, respectively), this gives a value of $\beta = -0.895$ for Reference, indicating it was the lowest-performing concept overall. There was a significant negative interaction between Emotion and English-at-school ($\beta = -0.018$, $SE = 0.004$, $z = -4.33$, $p < .001$), showing that the positive effect of L2-exposure was attenuated for Emotion relative to the overall pattern. No other concept-by-exposure interactions were significant. Additionally, L1-vocabulary ($\beta = 0.240$, $SE = 0.067$, $z = 3.56$, $p < .001$) and NVR ($\beta = 0.474$, $SE = 0.069$, $z = 6.91$, $p < .001$) were significant predictors of ToM accuracy. English-exposure outside school also

[9]Concept-Level-ToM refers to those non-ceiling concepts included in the analyses. It is not modeled directly as a total score but reflects aggregated performance derived from Concept-Level accuracy analyses.

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

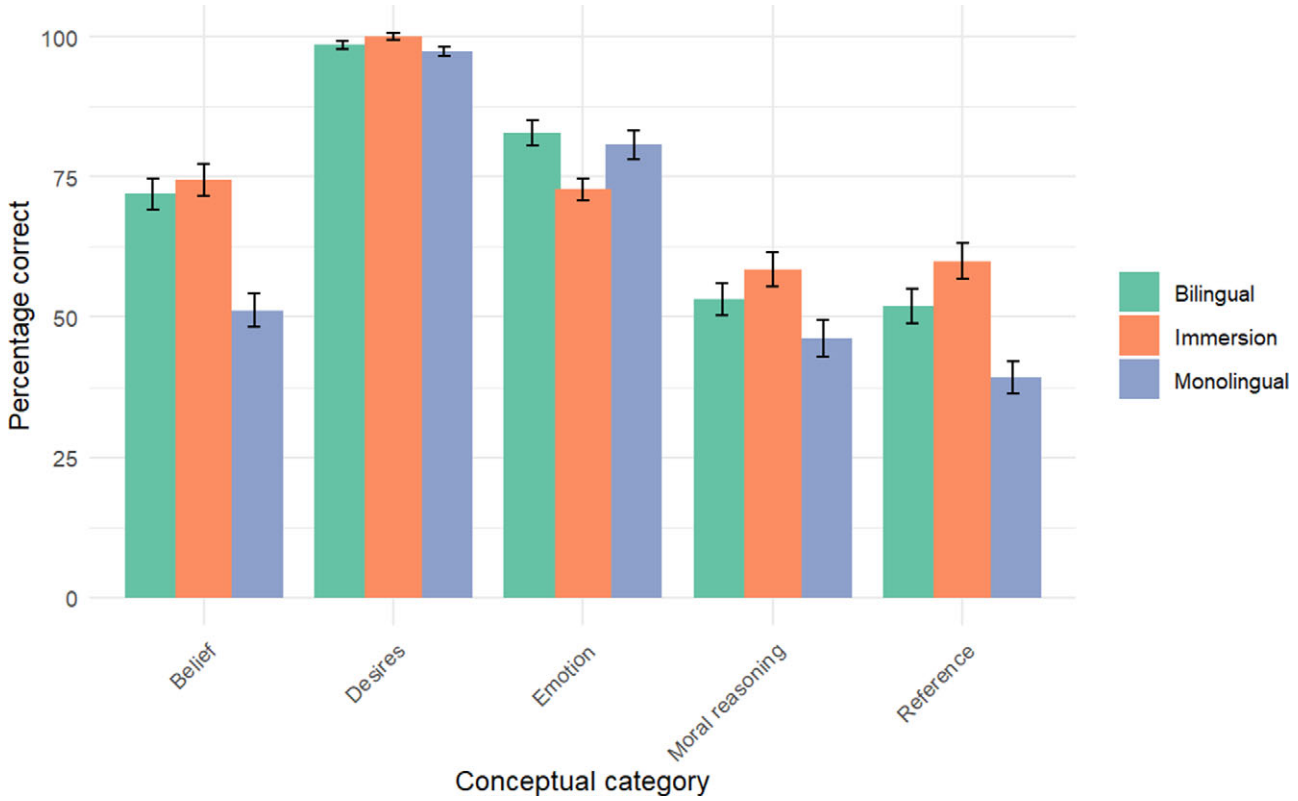

Figure 2. Accuracy percentages for each ToM concept by group at T0 (these values are intended for illustrative purposes only and should not be used to draw inferential conclusions).

contributed positively (β = 0.591, *SE* = 0.295, *z* = 2.01, *p* = .045). Conversely, L2-vocabulary was negatively associated with ToM (β = −0.196, *SE* = 0.079, *z* = −2.48, *p* = .013).

3.2. T0 versus T1

At T1, children across all three groups scored at/above the 95% ceiling threshold on Desires and Emotion, so at T1, these concepts were thus excluded from the main analyses. Figure 3 shows the accuracy percentages for each concept by group.

The model fitted to assess progress over time (T0 versus T1) on ToM (Belief, Reference, Moral-Reasoning) showed a significant effect of time (β = 1.001, *SE* = 0.096, *z* = 11.37, *p* < .001), indicating that all children's performance improved between T0 and T1. The model also revealed a significant main effect of English-at-school across Belief, Reference, and Moral-Reasoning (β = 0.012, *SE* = 0.003, *z* = 4.43, *p* < .001), indicating that greater English-at-school continued to enhance ToM. Using deviation coding, Concept-Level coefficients reflect how each concept's accuracy deviates from the overall mean. Belief contributed significantly more to overall ToM than the grand mean (β = 0.363, *SE* = 0.103, *z* = 3.51, *p* < .001), while the contribution of Moral-Reasoning did not differ significantly from it (β = −0.032, *SE* = 0.129, *z* = −0.24, *p* = .807). Despite Reference being the implicit baseline and therefore not directly estimated, its effect can be inferred as the inverse of the sum of the other two concepts ($\beta \approx -0.331$), illustrating that performance on Reference was lower than the grand mean. This suggests children performed best on Belief and less well on Reference and Moral-Reasoning. The interaction between concept and time is reflected in the individual concept-time coefficients: Moral-Reasoning showed significantly greater improvement longitudinally than the overall mean (β = 0.575, *SE* = 0.124, *z* = 4.65, *p* < .001), while Belief did not (β = 0.025, *SE* = 0.113, *z* = 0.22, *p* = .826) and Reference still less so. A significant three-way interaction between Moral-Reasoning, English-at-school, and time (β = 0.011, *SE* = 0.003, *z* = 4.14, *p* < .001) suggests that the positive effect of school L2-exposure on Moral-Reasoning became stronger over time. Conversely, a negative interaction between English-at-school and time (β = −0.008, *SE* = 0.002, *z* = −3.62, *p* < .001) indicated that while L2-exposure was beneficial overall, this effect slightly decreased between T0 and T1 (see Figure 4). Turning to individual-difference measures, L1-vocabulary (β = 0.207, *SE* = 0.045, *z* = 4.58, *p* < .001) and NVR (β = 0.223, *SE* = 0.040, *z* = 5.64, *p* < .001) were significant predictors of performance on these three cognitive concepts. In contrast, gender was not (*p* > .05). With respect to L2-vocabulary, although it was no longer negatively associated with ToM, as it had been at T0, it was also not a significant predictor of ToM at the year's end.

The final ToM analyses examined concepts introduced at T1 (Lies, Sarcasm) by fitting these concepts to the T1 data (see Supplementary Materials, Table S11). English-at-school did not significantly predict performance (β = −0.001, *SE* = 0.012, *z* = −0.02, *p* = .977), and children performed significantly worse on Sarcasm than Lies (β = −4.00, *SE* = 0.896, *z* = −4.45, *p* < .001). Children with higher NVR scores achieved better scores on these items (β = 0.354, *SE* = 0.121, *z* = 2.94, *p* = .003) as did older children (β = 0.679, *SE* = 0.266, *z* = 2.55, *p* = .011). While gender did not yield a significant main effect (*p* > .05), a significant interaction between concept and gender showed that girls outperformed boys on Sarcasm (β = −0.603, *SE* = 0.205, *z* = −3.08, *p* = 0.004).

3.3. Attention results

Our secondary question asked whether attention skills (selective, switching, response inhibition) contributed to ToM performance.

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

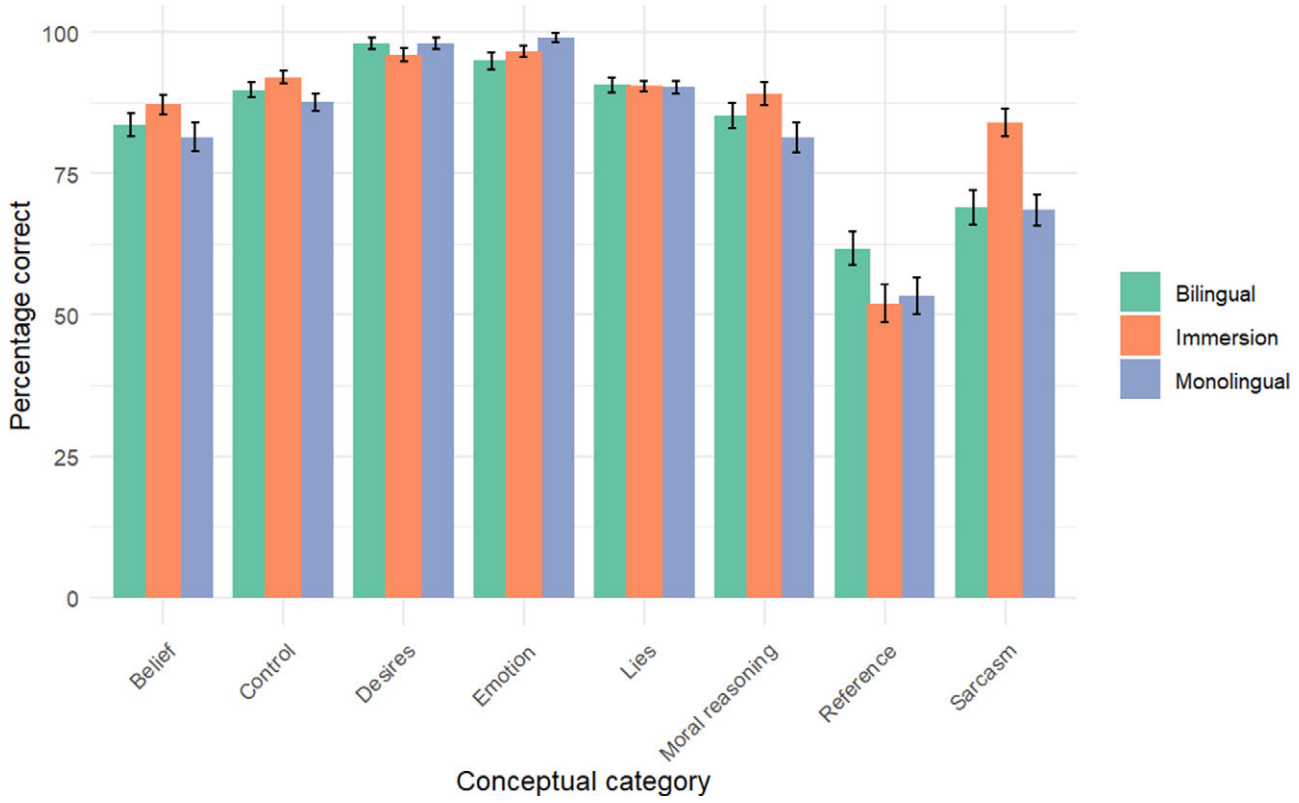

Figure 3. Accuracy percentages for each ToM concept by group at T1 (these values are intended for illustrative purposes only and should not be used to draw inferential conclusions).

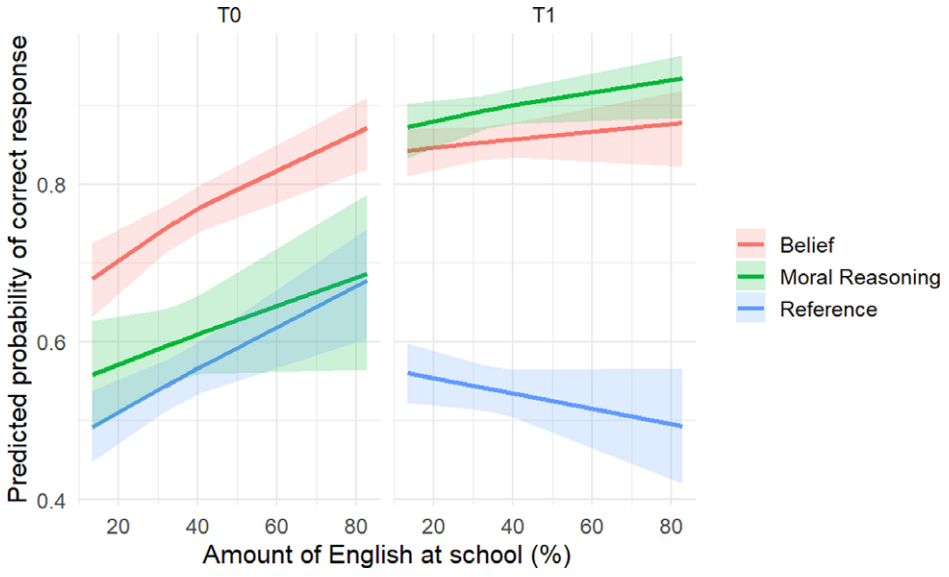

Figure 4. Interaction effect showing predicted probabilities of correct responses across Belief, Moral-Reasoning, and Reference as a function of English-at-school (%) by time point (T0 versus T1).

Thus, we extended our models by including attention measures as predictors at T0 and T1.[10]

First, we examined correlations among attention measures and found no evidence of problematic multicollinearity. The one strong correlation was between the two OppositeWorlds conditions ($r = 0.72$). This was expected given their shared task structure; however, both were retained because they represent distinct processing demands: the congruent condition requires task execution, while the incongruent one includes switching and inhibition. The full correlation matrix is in the Supplementary Materials (Table S13).

[10]Due to space, we only report attention skills and ToM results. See Supplementary Materials for full models with covariates.

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

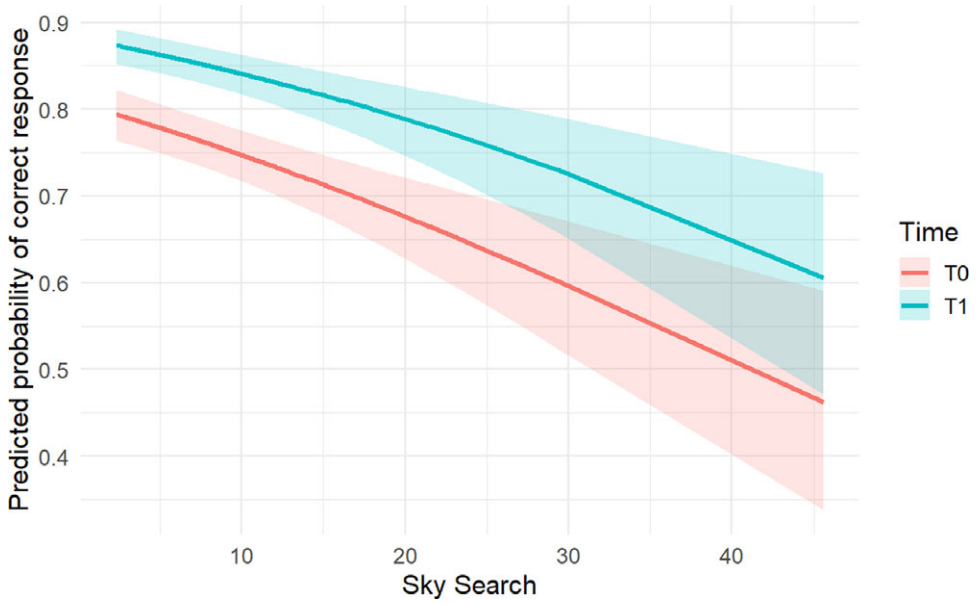

Figure 5. Interaction effect showing predicted probabilities of correct responses on ToM as a function of SkySearch performance by time point (T0 versus T1).

We extended our baseline model to explore the impact of attention measures on ToM at T0 (Desires excluded) and found that only selective attention and switching were significant contributors to ToM scores: children with higher SkySearch scores ($\beta = -0.028$, $SE = 0.010$, $z = -2.83$, $p = .004$) and greater CreatureCounting accuracy ($\beta = 0.073$, $SE = 0.031$, $z = 2.37$, $p = .017$) demonstrated superior performance. Consistent with our baseline model, L1-vocabulary ($\beta = 0.173$, $SE = 0.067$, $z = 2.56$, $p = .010$), NVR ($\beta = 0.393$, $SE = 0.069$, $z = 5.66$, $p < .001$), and English-at-school ($\beta = 0.010$, $SE = 0.003$, z = 2.87, $p = .004$) remained significant predictors of ToM, and BPVS was again negatively associated with ToM accuracy ($\beta = -0.205$, $SE = 0.079$, $z = -2.60$, $p = .009$).

We then examined progress between T0 and T1. SkySearch remained a significant predictor across the Belief, Moral-Reasoning, and Reference concepts ($\beta = -0.038$, $SE = 0.007$, $z = -5.19$, $p < .001$), an effect which strengthened over time ($\beta = 0.033$, $SE = 0.013$, $z = 2.44$, $p = .015$) (see Figure 5). Walk/Don't Walk was a significant positive predictor ($\beta = 0.025$, $SE = 0.012$, $z = 2.04$, $p = .041$), suggesting a small but reliable role of response inhibition in ToM performance, and CreatureCounting accuracy showed a positive trend toward significance ($\beta = 0.044$, $SE = 0.023$, $z = 1.90$, $p = .058$). However, unlike SkySearch, the interaction of time with CreatureCounting and Walk/Don't Walk did not reach significance (see Supplementary Materials, Table S9). These findings indicate that at T1, where not just Desires but also Emotion had been excluded from the model due to ceiling effects, selective attention, response inhibition, and switching all contributed toward ToM performance but that this was most marked for selective attention.

Finally, we ran models for the two solely T1 categories (Lies, Sarcasm). No attention measures reached significance for either of them ($p > .05$).

4. Discussion

Our investigation asked whether L2-exposure impacted ToM in Spanish children educated but not raised bilingually. Restricting our population in this way, we reduced the effects of multilingual/multicultural home on ToM (Tiv et al., 2021), and by including English-at-school as a continuous variable, we established firmer control over the timing and degree of L2-exposure. Children participated at the start and end of their first PE year and included monolinguals, bilinguals, and immersion children. Our foremost aim was to assess the effect of English-at-school on Concept-Level-ToM longitudinally so that we could ascertain whether the effect of English-at-school differed across specific ToM concepts. The second was to examine whether EF predicted ToM beyond that of L2-exposure, and the third was to monitor individual-difference contributions to ToM (Navarro et al., 2022).

To summarize, over time, children improved on all components and the order of difficulty largely reflected that reported in previous literature. English-at-school predicted Concept-Level-ToM at both time points, meaning that children with greater L2-exposure started and finished ahead. However, their initially steep improvement flattened, while children with lower exposure made more gains. With respect to attention, at T0, only selective attention and switching scores predicted Concept-Level-ToM, whereas at T1, all three attention measures did. Yet, the effect was most consistent for selective attention, whose effect also strengthened over time. Neither English-at-school nor any of the three attention measures was linked to Lies and Sarcasm. The key individual-difference predictors for the conflated Reference, Belief, and Moral-Reasoning ToM scores were NVR and L1-vocabulary. Gender was not a predictor for these cognitive ToM tasks. However, girls outperformed boys on Sarcasm, the one conative task. Contrary to expectations, higher L2-vocabulary scores did not predict ToM outcomes at either time points.

4.1. Background variables and ToM-component difficulty

At T0, there were no group differences with respect to age, NVR, WM, or L1-vocabulary, but children with more English-at-school had higher L2-vocabulary. Eighty-five percent of children had attended preschool at their respective school so had already experienced different amounts of English prior to PE, demonstrating that

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

even at preschool, where English activities are less structured, degree of L2-exposure yields visible linguistic consequences.

Our analyses indicated the following difficulty ranking of concepts at T0: Desires>Emotion>Belief>Moral-Reasoning>Reference (see Supplementary Materials, Figures S2 and S3). Emotion, classed as affective, proved easier than all but one of the concepts classed as cognitive, in line with Shamay-Tsoory et al. (2007) which reported better performance on affective than cognitive tasks albeit on tasks designed for older children and adults. Our finding that Desires was easier than Belief, which was easier than Reference, accords with the proposed order of difficulty described in Wellman and Liu (2004) and Peterson and Wellman (2019), although Emotion, which our populations found easier than Belief, is at odds with the latter study's findings for typical monolingual children and deaf children born to hearing parents. This contrast should be interpreted cautiously, however, and in relation to the different levels of difficulty between these test batteries.

The ranking found at T0 remained at T1, although the gap between Moral-Reasoning and Reference increased. The two new concepts were interspersed on this scale. Lies achieved near-ceiling scores, and Sarcasm, although more difficult than Belief, was understood more easily than Reference, which remained the most challenging. Difficulties with Sarcasm were expected given its conative classification (Dennis et al., 2013), and Lies can be explained by these particular task requirements. Children only had to distinguish a fact from a false state-of-affairs, not infer an intention behind the deception, likening the task to an easier subcomponent of FB (Astington & Gopnik, 1991). Their excellent performance makes sense given their age at T1. Sarcasm, however, where children needed to understand that a protagonist was attempting to exert influence over another and that an intended meaning of a statement could conflict with the literal meanings of the words within it, proved one of the most difficult, as per Peterson et al. (2012). With respect to Reference, we think a particular conceptualization of Reference in one vignette answers for this category's overall scores. Whereas for most items there was a straightforward spatial relationship to be navigated, one item incorporated an implicit size comparison, leading children to focus on an incorrect reference point from the outset. Their mistakes, therefore, were less about Reference than on how to draw comparisons between sizes when key contextual clues are missed (see 'Task difficulty' in Supplementary Materials). With these items – 2 of 6 – omitted, the children's score improved substantially, suggesting a better understanding of Reference than initially illustrated, placing it in line with Sarcasm.

4.2. English at school and ToM

The fact that English-at-school exhibited an effect at both time points suggests that the spur to ToM provided by educational bilingualism takes effect quickly. Our results accord with Agostini et al. (2025), whose L2-learners, after 1 year of immersion, also outperformed monolinguals on perspective taking. Relatedly, Listanti et al. (2023) found that even by increasing reading in children's heritage language, first- and second-order FB improved. Taken together, these results suggest that the L2-threshold for influence on one aspect of ToM can be quite low if children are not below 4 years (Cheung et al., 2010). Practically, our results suggest that at this early stage, it is the bilingual experience itself that has positive consequences for ToM rather than a prescribed level of L2-attainment, meaning all children could profit from early access to an L2. Further support for this proposition is that children's L2-vocabulary scores did not predict ToM at either time points. The lack of an L2-vocabulary effect is consistent with the cognitive exercise and/or socio-cultural experience of bilingualism being the relevant ToM-boosting conduits as opposed to L2-fluency. Also of note is the negative interaction between English-at-school and time at T1, which indicates that although children with greater L2-exposure continued to score higher than those with less exposure, the magnitude of the advantage diminished. Such a pattern suggests that it was the lower-exposure children who made more gains during this period. Continued tracking will indicate if this reflects an attenuation of the bilingual children's progress or whether further boosts in the bilinguals' performance materialize over time. The results are all the more interesting given current debates with respect to what causes the bilingualism effect, in particular, Paap (2019), who suggests that the initial spur to EF provided by bilingualism is short-lived and due to the sudden increase in intentional cognitive effort at the incipient stage of (language) learning. Given the purported partial overlap in requirements underlying cognitive-ToM and attention tasks (Wellman et al., 2001), one might expect L2-exposure to impact cognitive-ToM and EF progression similarly. However, Chamorro et al. (2025), who examined attention in the same children reported here, found that the positive effects of L2-exposure on selective attention, switching, and response inhibition increased longitudinally. Continued monitoring of these children's trajectory with respect to both EF and all three ToM components will clarify which, if any, are similarly affected by L2-exposure. Returning to the contribution of English-at-school on ToM in the current study (which excluded Desires at T0 and Desires and Emotion at T1), an L2-exposure effect was present at both time points on the remaining cognitive tasks. This is consistent with Yu et al. (2021), who reported a bilingual advantage for ToM in 16 of the 21 studies reviewed, where the majority used FB, a quintessentially cognitive task.

An interesting question is why English-at-school only interacted negatively with Emotion. However, this task, classed as affective, would draw on further competencies to those required for purely cognitive tasks, so skills honed by bilingualism that influence these might not generalize to affective ones (Kalbe et al., 2010; Yott & Poulin-Dubois, 2016). Indeed, studies on adults with clinical profiles have found behavioral and neurological dissociations between affective- and cognitive-ToM tasks (Healey & Grossman, 2018; Shamay-Tsoory et al., 2007). Recall also Cassetta et al. (2018), who reported that inhibitory control and switching predicted performance on cognitive-ToM tasks but not affective ones. It might be that alternative consequences of bilingualism, such as increased sociocultural awareness – a measure we did not include – are more pertinent for affective aspects of ToM development (Cheung et al., 2010). If true, we would not necessarily expect our monolingually raised bilinguals to exhibit an early advantage for Emotion since their access to the broader cultural experience that might accelerate sociocultural awareness was reduced (Han & Lee, 2013; Tiv et al., 2021). Further monitoring over time with more challenging Emotion tasks would tell.

English-at-school did not predict Lies/Sarcasm scores either. For Lies, all children achieved over 90% accuracy, so the lack of effect is likely due to this near-ceiling performance. However, Sarcasm performance might relate to the further social and empathetic components inherent to this conative task, depending as it does on engagement with social norms and feelings. Children had to interpret superficially contradictory statements, calling upon sophisticated social reasoning skills to forge the link between seemingly antithetical displays of behavior and protagonists'

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

intended meanings. Enhancement of such social reasoning skills might again not materialize in children not brought up multiculturally: again, further monitoring over time will tell.

4.3. Attention and ToM

The influence of attention grew with time. Whereas at T0, only selective attention and switching predicted ToM, at T1, selective attention, switching, and inhibition did. However, it was only selective attention whose influence strengthened over time.

Selective attention requires participants to attend to relevant stimuli and to ignore distractors. This process is analogous to that needed to choose between correct and incorrect beliefs/visual perspectives and between physical appearances and reality. In all these cognitive tasks, interfering or distracting information must be suppressed in favor of less salient yet correct information. On this basis, one might expect that if L2-exposure boosts selective attention (Adesope et al., 2010), the influence of selective attention would be similarly visible on the prototypically cognitive tasks as observed here.

Aspects of switching were also significant for ToM, and the links between switching and ToM per se ties in with previous literature. The cognitive action of language switching is arguably akin to the practice of adapting quickly to changing task requirements (as in DCCS and CreatureCounting and OppositeWorlds), a similarity which would answer for the frequently found advantage for bilinguals on cognitive flexibility (Bialystok, 2010). What all ToM tasks share is that participants must alternate between different viewpoints, be it a belief, a feeling, or a visual perspective, which points to an overlap in this cognitive action and that underpinning switching tasks. The correspondence we found, therefore, between switching and ToM in bilinguals might be anticipated: if early bilinguals surpass monolinguals on switching, this should translate to some of the mental operations engaged during ToM tasks (Bialystok & Martin, 2004; Prior & McWhinney, 2010). Indeed, Austin et al. (2014) found that switching predicted ToM in monolingual children, and Buac and Kaushanskaya (2020) reported it as integral to bilingual/immersion children's FB scores (see also Bialystok & Viswanathan, 2009; Carlson & Meltzoff, 2008). Where our results offer pause for thought, however, is that educational bilingualism does not provide the same opportunities for language switching that sequential or simultaneous bilinguals enjoy, and it is such an environment that has been argued to hone this skill (Bialystok, 2010). The current bilingual children exhibited heightened switching skills, as reported in Chamorro et al. (2025), in conjunction with a positive association between switching and ToM, despite not having the environment most conducive to its enhancement.

Our results for response inhibition are not entirely in sync with what one might expect if inhibition abilities underpin ToM because this variable only exerted an influence at T1. The positive impact of bilingualism on inhibition is widely supported (Barac et al., 2014; Bialystok & Martin, 2004; but see also Bialystok & Viswanathan, 2009; Carlson & Meltzoff, 2008; Duñabeitia et al., 2014), and indeed Chamorro et al. (2025) found enhanced inhibition in these very bilinguals. Despite this, inhibition's contribution to the children's ToM scores at this incipient stage was less central. Recall, too, that Huang et al. (2023) reported no relation between inhibition and ToM scores, unlike switching. Still more recently, Agostini et al. (2024) reported that inhibition had not been integral to a bilingual advantage they found for referential perspective taking. These discrepant results motivate further exploration into why response inhibition showed a weaker association with ToM at this stage than did switching and selective attention.

4.4. Individual differences and ToM

Of all the covariates, the three that made notable contributions were NVR, L1-vocabulary, and gender. L2-vocabulary, however, was conspicuous for its absence. With respect to gender, this variable was of relevance to the Sarcasm task, where girls outperformed boys. Performance on this conative task, which calls on social reasoning and empathic skills for its successful execution, was also unaffected by any of the attention measures or English-at-school. In this respect, it behaved similarly to Emotions, which interacted negatively with English-at-school. Thus, results for the two tasks that are not purely cognitive diverged from those that are: proportion of English-at-school was not relevant to performance at this point of exposure. A future study might explore if this dichotomy between cognitive tasks on the one hand and conative on the other remains over time and extends to a broader range of bilingual contexts.

With respect to L1-vocabulary, our results support the monolingual literature, which has shown strong vocabulary skills in monolinguals predict ToM (de Villiers & de Villiers, 2014). Our tasks included extensive explanatory aspects, where children justified decisions. Better justification of decisions is linked to strong language skills (Białecka et al., 2024), so the relevance of L1-vocabulary we found throughout makes sense conceptually. The lack of a relation between L2-vocabulary and ToM, however, is unexpected, and contrasts with Chamorro et al. (2025), where this variable predicted performance on selective attention, switching, and inhibition. The apparent independence of ToM from L2-vocabulary at this stage is something to be monitored over time as this also relates to the question over whether the same factors that promote EF also impact on ToM. Lastly, the bearing of NVR on every ToM component indicates the role played by fluid intelligence in these tasks all of which encompass cognition.

5. Conclusions

Our study has shown that in educational bilingualism, the impact of L2-exposure on ToM is visible immediately and not dependent on L2-vocabulary attainment. This means educators need not worry that these mentalizing benefits are restricted to children who can reach a certain L2-threshold. Aspects of attention also predicted ToM but not uniformly so. The influence of selective attention was the most consistent and its effects strengthened over time. We have argued that the characteristics underlying cognitive aspects of ToM tasks map closely to those underlying attention tasks, already shown by many to be enhanced in bilinguals. However, Emotion, an affective task, and Sarcasm, a conative one, stood out in that they were not affected by L2-exposure and girls did better than boys on Sarcasm. Further explorations of these components, with a harder Emotions task than the current one, could explore in more depth whether and how L2-exposure and gender interact with cognitive-, affective-, and conative-ToM distinctions. We have also suggested that the qualitatively different experience of educational bilingualism, where children's exposure to multilingual/cultural experiences is reduced, might affect when and if bilingual gains on affective and conative skills surface. The fact that girls surpassed boys on conative tasks but not cognitive ones, together with the lack of any large systematic gender divide on attention tasks (Grissom & Reyes, 2019), is a further reason to conclude that although ToM and EF share some underpinnings, they should not be conflated. In particular, the effects of the characteristics of ToM components that incorporate sentiments and heightened social awareness are

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

unearthed by using concept-level rather than composite scoring. Finally, the consistent influence of NVR and L1-vocabulary on performance across the board reminds us of how integral these variables are to ToM development, irrespective of bilingual exposure.

Supplementary Material. The supplementary material for this article can be found at http://doi.org/10.1017/S1366728926101229.

Data availability statement. The R script used for the analyses is available via the embedded link. The data set forms part of an ongoing longitudinal project. As data collection is still underway, we are unable to share the data set at this time to protect participant confidentiality and preserve the integrity of future analyses. The data set will be made available upon completion of the study.

Acknowledgments. This project was funded by the Comunidad de Madrid (Atracción de Talento Investigador, T1/HUM-19952). We would also like to thank the participating schools, parents, and children.

Competing interests. The authors declare no competing interests.

References

Adesope, O. O., **Lavin, T.**, **Thompson, T.**, & **Ungerleider, C.** (2010). A systematic review and meta-analysis of the cognitive correlates of bilingualism. *Review of Educational Research*, **80**, 207–245.

Agostini, V., **Apperly, I.**, & **Krott, A.** (2024, January 9). *Inhibitory control or social skills? Two accounts for a bilingual advantage*. https://doi.org/10.31234/osf.io/b3mht.

Agostini, V., **Apperly, I.**, & **Krott, A.** (2025). Greater sensitivity to communication partners' perspectives in children learning a second language at school. *Bilingualism: Language and Cognition*, **10**, 1–15. https://doi.org/10.1017/S1366728925000069.

Antón, E., **Carreiras, M.**, & **Duñabeitia, J. A.** (2019). The impact of bilingualism on executive functions and working memory in young adults. *PLoS One*, **14**(2). https://doi.org/10.1371/journal.pone.0206770.

Apperly, I. (2011). *Mindreaders: The cognitive basis of "theory of mind*. Psychology Press.

Astington, J. W., & **Gopnik, A.** (1991). Theoretical explanations of children's understanding of the mind. *British Journal of Developmental Psychology*, **9**, 7–31.

Austin, G., **Groppe, K.**, & **Elsner, B.** (2014). The reciprocal relationship between executive function and theory of mind in middle childhood: A 1-year longitudinal perspective. *Frontiers in Psychology*, **5**, 655. https://doi.org/10.3389/fpsyg.2014.00655.

Ausubel, D. P., **Sullivan, E. V.**, & **Ives, S. W.** (1980). *Theory and problems of child development* (3rd ed.). Grune and Stratton.

Barac, R., **Bialystok, E.**, **Castro, D. C.**, & **Sanchez, M.** (2014). The cognitive development of young dual language learners: A critical review. *Early Childhood Research Quarterly*, **29**(4), 699–714. https://doi.org/10.1016/j.ecresq.2014.02.003

Baron-Cohen, S., **O'Riordan, M.**, **Stone, V.**, **Jones, R.**, & **Plaisted, K.** (1999). Recognition of faux pas by normally developing children and children with Asperger syndrome or high-functioning autism. *Journal of Autism and Developmental Disorders*, **29**(5), 407–418. https://doi.org/10.1023/A:1023035012436.

Barr, D. J., **Levy, R.**, **Scheepers, C.**, & **Tily, H. J.** (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, **68**(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001.

Białecka, M., **Wodniecka, Z.**, **Muszyńska, K.**, **Szpak, M.**, & **Haman, E.** (2024). Both L1 and L2 proficiency impact ToM reasoning in children aged 4 to 6. Painting a more nuanced picture of the relation between bilingualism and ToM. *Bilingualism: Language and Cognition*, **27**(3), 400–418. https://doi.org/10.1017/S1366728923000652.

Bialystok, E. (2010). Global-local and trail-making tasks by monolingual and bilingual children: Beyond inhibition. *Developmental Psychology*, **46**(1), 93–105. https://doi.org/10.1037/a0015466.

Bialystok, E., & **Craik, F. I. M.** (2010). Cognitive and linguistic processing in the bilingual mind. *Current Directions in Psychological Science*, **19**(1), 19–23. https://doi.org/10.1177/0963721409358571.

Bialystok, E., & **Majumder, S.** (1998). The relationship between bilingualism and the development of cognitive processes in problem solving. *Applied PsychoLinguistics*, **19**(1), 69–85. https://doi.org/10.1017/S0142716400010584.

Bialystok, E., & **Martin, M. M.** (2004). Attention and inhibition in bilingual children: Evidence from the dimensional change card sort task. *Developmental Science*, **7**(3), 325–339. https://doi.org/10.1111/j.1467-7687.2004.00351.x.

Bialystok, E., & **Senman, L.** (2004). Executive processes in appearance-reality tasks: The role of inhibition of attention and symbolic representation. *Child Development*, **75**, 562–579. https://doi.org/10.1111/j.1467-8624.2004.00693.x.

Bialystok, E., & **Viswanathan, M.** (2009). Components of executive control with advantages for bilingual children in two cultures. *Cognition*, **112**(3), 494–500. https://doi.org/10.1016/j.cognition.2009.06.014.

Branzi, F. M., **Calabria, M.**, **Gade, M.**, **Fuentes, L. J.**, & **Costa, A.** (2016). On the bilingualism effect in task switching. *Bilingualism: Language and Cognition*, **21**, 1–14. https://doi.org/10.1017/S136672891600119X.

Buac, M., & **Kaushanskaya, M.** (2020). Predictors of theory of mind performance in bilingual and monolingual children. *International Journal of Bilingualism*, **24**(2), 339–359. https://doi.org/10.1177/1367006919826866.

Bukhalenkova, D., **Veraksa, A.**, **Gavrilova, M.**, & **Kartushina, N.** (2022). Emotion understanding in bilingual preschoolers. *Behavioral Sciences*, **12**(4), 115. https://doi.org/10.3390/bs12040115.

Cape, R., **Vega-Mendoza, M.**, **Bak, T. H.**, & **Sorace, A.** (2018). Cognitive effects of Gaelic medium education on primary school children in Scotland. *International Journal of Bilingual Education and Bilingualism*, 1–20. https://doi.org/10.1080/13670050.2018.1543648.

Carlson, S. M., & **Meltzoff, A. N.** (2008). Bilingual experience and executive functioning in young children. *Developmental Science*, **11**, 279–295. https://doi.org/10.1111/j.1467-7687.2008.00675.x.

Carlson, S. M., **Moses, L. J.**, & **Breton, C.** (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and WM. *Infant and Child Development*, **11**, 73–92. https://doi.org/10.1002/icd.298.

Cassetta, B. D., **Pexman, P. M.**, & **Goghari, V. M.** (2018). Cognitive and affective theory of mind and relations with executive functioning in middle childhood. *Merrill-Palmer Quarterly*, **64**(4), 514–538. https://doi.org/10.13110/merrpalmquar1982.64.4.0514.

Chamorro, G., **de la Viña, I.**, & **Janke, V.** (2025). The effects of L2-exposure at school on the cognitive development of children from monolingual backgrounds: A longitudinal study. *Bilingualism, Language & Cognition*, **3**, 1–15. https://doi.org/10.1017/S1366728925100448.

Chamorro, G., & **Janke, V.** (2020). Investigating the bilingual advantage: the impact of L2 exposure on the social and cognitive skills of monolingually-raised children in bilingual education. *International Journal of Bilingual Education and Bilingualism*, **25**(5), 1765–1781. https://doi.org/10.1080/13670050.2020.1799323

Chamorro, G., & **Janke, V.** (2023). Educational bilingualism: Reflections on a longitudinal study of children's cognitive and linguistic development. *Ampersand*, **10**, 100115. https://doi.org/10.1016/j.amper.2023.100115

Cheung, H., **Mak, W. Y.**, **Luo, X.**, & **Xiao, W.** (2010). Sociolinguistic awareness and false belief in young Cantonese learners of English. *Journal of Experimental Child Psychology*, **107**(2), 188–194. https://doi.org/10.1016/j.jecp.2010.05.001.

Costa, A., **Hernández, M.**, **Costa-Faidella, J.**, & **Sebastián-Gallés, N.** (2009). On the bilingual advantage in conflict processing: Now you see it, now you don't. *Cognition*, **113**, 135–149. https://doi.org/10.1016/j.cognition.2009.08.001.

Costa, A., **Hernández, M.**, & **Sebastián-Gallés, N.** (2008). Bilingualism aids conflict resolution: Evidence from the ANT task. *Cognition*, **106**(1), 59–86. https://doi.org/10.1016/j.cognition.2006.12.013.

Duñabeitia, J. A., **Hernández, J. A.**, **Antón, E.**, **Macizo, P.**, **Estévez, A.**, **Fuentes, L. J.**, & **Carreiras, M.** (2014). The inhibitory advantage in bilingual children revisited: Myth or reality?. *Experimental Psychology*, **61**(3), 234–251. https://doi.org/10.1027/1618-3169/a000243

de Villiers, J. G., & **de Villiers, P. A.** (2014). The role of language in theory of mind development. *Topics in Language Disorders*, **34**(4), 313–328. https://doi.org/10.1097/TLD.0000000000000037.

Dennis, M., **Simic, N.**, **Bigler, E. D.**, **Abildskov, T.**, **Agostino, A.**, **Taylor, H. G.**, **Rubin, K.**, **Vannatta, K.**, **Gerhardt, C. A.**, **Stancin, T.**, & **Yeates, K. O.** (2013). Cognitive, affective, and conative theory of mind (ToM) in children with traumatic brain injury. *Developmental Cognitive Neuroscience*, **5**, 25–39. https://doi.org/10.1016/j.dcn.2012.11.006.

Devine, R. T., & **Hughes, C.** (2014). Relations between false-belief understanding and executive function in early childhood: A meta-analysis. *Child Development*, **85**, 1777–1794. https://doi.org/10.1111/cdev.12237.

Diaz, V., & **Farrar, M. J.** (2017). The missing explanation of the false-belief advantage in bilingual children: A longitudinal study. *Developmental Science*, **21**(4), e12594. https://doi.org/10.1111/desc.12594.

Diaz, V., & **Farrar, M. J.** (2018). Do bilingual and monolingual preschoolers acquire false belief understanding similarly? The role of executive functioning and language. *First Language*, **38**(4), 382–398. https://doi.org/10.1177/0142723717752741.

Dick, A. S., **Garcia, N. L.**, **Pruden, S. M.**, **Thompson, W. K.**, **Hawes, S. W.**, **Sutherland, M. T.**, **Riedel, M. C.**, **Laird, A. R.**, & **Gonzalez, R.** (2019). No evidence for a bilingual executive function advantage in the ABCD study. *Nature Human Behaviour*, **3**, 692–701. https://doi.org/10.1038/s41562-019-0609-3.

Doenyas, C., **Yavuz, H. M.**, & **Selcuk, B.** (2018). Not just a sum of its parts: How tasks of the theory of mind scale relate to executive function across time. *Journal of Experimental Child Psychology*, **166**, 485–501. https://doi.org/10.1016/j.jecp.2017.09.014

Dunn, L. M., **Dunn, D. M.**, & **Styles, B.** (2009). *The British picture vocabulary scale - third edition*. GL Assessment.

Dunn, L. M., **Dunn, L. M.**, & **Arribas, D.** (2006). *PPVT-III Peabody: Test de vocabulario en imágenes: Manual*. TEA Ediciones.

Engel de Abreu, P. M. J., **Cruz-Santos, A.**, **Tourinho, C. J.**, **Martin, R.**, & **Bialystok, E.** (2012). Bilingualism enriches the poor: Enhanced cognitive control in low-income minority children. *Psychological Science*, **23**, 1364. https://doi.org/10.1177/0956797612443836.

Farhadian, M., **Abdullah, R.**, **Mansor, M.**, **Redzuan, M. A.**, **Kumar, V.**, & **Gazanizad, N.** (2010). Theory of mind, birth order, and siblings among preschool children. *American Journal of Scientific Research*, **7**(3), 25–35.

Feng, J., **Cho, S.**, & **Luk, G.** (2023). Assessing theory of mind in bilinguals: A scoping review on tasks and study designs. *Bilingualism: Language and Cognition*, **27**(4), 531–545. https://doi.org/10.1017/s1366728923000585

Flavell, J. H., **Flavell, E. R.**, & **Green, F. L.** (1983). Development of the appearance–reality distinction. *Cognitive Psychology*, **15**, 95–120. https://doi.org/10.1016/0010-0285(83)90005-1.

Genesee, F. (2004). What do we know about bilingual education for majority language students? In Bhatia, T. K. & Ritchie, W. (Eds.), *Handbook of bilingualism and multiculturalism* (pp. 547–576). Blackwell.

Goetz, P. J. (2003). The effects of bilingualism on theory of mind development. *Bilingualism: Language and Cognition*, **6**, 1–15. https://doi.org/10.1017/S1366728903001007.

Greenberg, A., **Bellana, B.**, & **Bialystok, E.** (2013). Perspective-taking ability in bilingual children: Extending advantages in executive control to spatial reasoning. *Cognitive Development*, **28**, 41–50. https://doi.org/10.1016/j.cogdev.2012.10.002.

Grissom, N. M., & **Reyes, T. M.** (2019). Let's call the whole thing off: Evaluating gender and sex differences in executive function. *Neuropsychopharmacology*, **44**(1), 86–96. https://10.1038/s41386-018-0179-5.

Grote, K. S., **Scott, R. M.**, & **Gilger, J.** (2021). Bilingual advantages in executive functioning: Evidence from a low-income sample. *First Language*, **41**(6), 677–700. https://doi.org/10.1177/01427237211024220

Hakuta, K. (1986). *Mirror of language: The debate on bilingualism*. Basic Books.

Han, S., & **Lee, K.** (2013). Cognitive and affective perspective-taking ability of young bilinguals in South Korea. *Child Studies in Diverse Contexts*, **3**(1), 69–80. https://doi.org/10.5723/csdc.2013.3.1.069.

Hartig, F. (2022). *DHARMa: Residual diagnostics for hierarchical (multi-level / mixed) regression models. R package version 0.4.7*. CRAN.

Healey, M. L., & **Grossman, M.** (2018). Cognitive and affective perspective-taking: Evidence for shared and dissociable anatomical substrates. *Frontiers in Neurology*, **9**, 491. https://doi.org/10.3389/fneur.2018.00491.

Hermanto, N., **Moreno, S.**, & **Bialystok, E.** (2012). Linguistic and metalinguistic outcomes of intense immersion education: How bilingual? *International Journal of Bilingual Education and Bilingualism*, **15**(2), 131–145. https://doi.org/10.1080/13670050.2011.652591.

Hernández, M., **Martin, C. D.**, **Barceló, F.**, & **Costa, A.** (2013). Where is the bilingual advantage in task-switching? *Journal of Memory and Language*, **69**(3), 257–276. https://doi.org/10.1016/j.jml.2013.06.004.

Huang, R., **Baker, E. R.**, & **Wang, T.** (2023). Early bilingualism enhances theory of mind in children from low-income households via executive function skills. *Cognitive Development*, **68**, 101389. https://doi.org/10.1016/j.cogdev.2023.101389.

Hughes, C., & **Devine, R.** (2019). Learning to read minds: A synthesis of social and cognitive perspective. In Whitebread, D., Grau, V., Kumpulainen, K., McClelland, M. M., Perry, N. E., & Pino-Pasternak, D. (Eds.), *The SAGE handbook of developmental psychology and early childhood education* (pp. 169–184). SAGE. https://doi.org/10.4135/9781526470393.

Hughes, C., **Jaffee, S. R.**, **Happé, F.**, **Taylor, A.**, **Caspi, A.**, & **Moffitt, T. E.** (2005). Origins of individual differences in theory of mind: From nature to nurture? *Child Development*, **76**(2), 356–370. https://doi.org/10.1111/j.1467-8624.2005.00850_a.x.

Kalashnikova, M., & **Mattock, K.** (2014). Maturation of executive functioning skills in early sequential bilingualism. *International Journal of Bilingual Education and Bilingualism*, **17**(1), 111–123. https://doi.org/10.1080/13670050.2012.746284.

Kalbe, E., **Schlegel, M.**, **Sack, A. T.**, **Nowak, D. A.**, **Dafotakis, M.**, **Bangard, C.**, **Shah, N. J.**, **Fink, G. R.**, & **Kessler, J.** (2010). Dissociating cognitive from affective theory of mind: A TMS study. *Cortex*, **46**, 769–780. https://doi.org/10.1016/j.cortex.2009.07.010.

Kovács, A. M. (2009). Early bilingualism enhances mechanisms of false-belief reasoning. *Developmental Science*, **12**, 48–54. https://doi.org/10.1111/j.1467-7687.2008.00742.x.

Lakens, D. (2022). Sample size justification. *Collabra: Psychology*, **8**(1), 33267. https://doi.org/10.1525/collabra.33267.

Liddle, B., & **Nettle, D.** (2006). Higher-order theory of mind and social competence in school-age children. *Journal of Cultural and Evolutionary Psychology*, **4**(3–4), 231–244. https://doi.org/10.1556/JCEP.4.2006.3-4.3.

Listanti, A., **Torregrossa, J.**, **Eisenbeiß, S.**, & **Bongartz, C.** (2023). Home literacy exposure in the heritage language enhances theory-of-mind development: A study on Greek-Italian bilingual children. In Proceedings of the 47thAnnual Boston University conference on language development *(BUCLD 47)*. Cascadilla Press.

Manly, T., **Robertson, I. H.**, **Anderson, V.**, & **Nimmo-Smith, I.** (1999). *TEA-Ch: The test of everyday attention for children manual*. Thames Valley Test Company Limited.

Marcovitch, S., **O'Brien, M.**, **Calkins, S. D.**, **Leerkes, E. M.**, **Weaver, J. M.**, & **Levine, D. W.** (2015). A longitudinal assessment of the relation between executive function and theory of mind at 3, 4, and 5 years. *Cognitive Development*, **33**, 40–55. https://doi.org/10.1016/j.cogdev.2014.07.001

Milligan, K., **Astington, J. W.**, & **Dack, L.** (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, **78**, 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x.

Navarro, E., **DeLuca, V.**, & **Rossi, E.** (2022). It takes a village: Using network science to identify the effect of individual differences in bilingual experience for theory of mind. *Brain Sciences*, **12**, 487. https://doi.org/10.3390/brainsci12040487.

Nguyen, T.-K., & **Astington, J. W.** (2014). Reassessing the bilingual advantage in theory of mind and its cognitive underpinnings. *Bilingualism: Language and Cognition*, **17**(2), 396–409. https://doi.org/10.1017/S1366728913000394.

Nichols, E. S., **Wild, C. J.**, **Stojanoski, B.**, **Battista, M. E.**, & **Owen, A. M.** (2020). Bilingualism affords no general cognitive advantages: A population study of executive function in 11,000 people. *Psychological Science*, **31**, 548–567. https://doi.org/10.1177/0956797620903113.

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press

Paap, K. R. (2019). The bilingual advantage debate: Quantity and quality of the evidence. In Schwieter, J. W. (Ed.), *The handbook of the neuroscience of multilingualism* (pp. 701–735). Wiley-Blackwell.

Paap, K. R., & **Greenberg, Z. I.** (2013). There is no coherent evidence for a bilingual advantage in executive processing. *Cognitive Psychology*, **66**(2), 232–258. https://doi.org/10.1016/j.cogpsych.2012.12.002.

Paap, K. R., **Johnson, H. A.**, & **Sawi, O.** (2015). Bilingual advantages in executive functioning either do not exist or are restricted to very specific and undetermined circumstances. *Cortex*, **69**, 265–278. https://doi.org/10.1016/j.cortex.2015.04.014.

Peterson, C. C., & **Wellman, H. M.** (2019). Longitudinal theory of mind (ToM) development from preschool to adolescence with and without ToM delay. *Child Development*, **90**(6), 1917–1934. https://doi.org/10.1111/cdev.13064.

Peterson, C. C., **Wellman, H. M.**, & **Slaughter, V.** (2012). The mind behind the message: Advancing theory-of-mind scales for typically developing children, and those with deafness, autism, or Asperger syndrome. *Child Development*, **83**(2), 469–485. https://doi.org/10.1111/j.1467-8624.2011.01728.x.

Plonsky, L., & **Ghanbar, H.** (2018). Multiple regression in L2 research: A methodological synthesis and guide to interpreting R2 values. *The Modern Language Journal*, **102**, 713–731. https://doi.org/10.1111/modl.12509.

Planckaert, N., **Duyck, W.**, & **Woumans, E.** (2023). Is there a cognitive advantage in inhibition and switching for bilingual children? A systematic review. *Frontiers in Psychology*, **14**. https://doi.org/10.3389/fpsyg.2023.1191816

Premack, D., & **Woodruff, G.** (1978). Does the chimpanzee have theory of mind? *Behavioral and Brain Sciences*, **1**(4), 515–526. https://doi.org/10.1017/S0140525X00076512.

Prior, A., & **Macwhinney, B.** (2010). A bilingual advantage in task switching. *Bilingualism: Language and Cognition*, **13**(2), 253–262. https://doi.org/10.1017/s1366728909990526

R Core Team. (2023). *R: A language and environment for statistical computing (version 4.2.1)*. R Foundation for Statistical Computing.

Raven, J. C., **Court, J. H.**, & **Raven, J.** (1998). *Progressive coloured matrices*. Oxford Psychologists Press.

Rubio-Fernández, P., & **Glucksberg, S.** (2012). Reasoning about other people's beliefs: Bilinguals have an advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **38**(1), 211–217. https://doi.org/10.1037/a0025162.

Schlaffke, L., **Lissek, S.**, **Lenz, M.**, **Juckel, G.**, **Schultz, T.**, **Schultz, T.**, **Tegenthoff, M.**, **Schmidt-Wilcke, T.**, & **Brüne, M.** (2015). Shared and nonshared neural networks of cognitive and affective theory-of-mind: A neuroimaging study using cartoon picture stories. *Human Brain Mapping*, **36**(1), 29–39. https://doi.org/10.1002/HBM.22610.

Schroeder, S. R. (2018). Do bilinguals have an advantage in theory of mind? A meta-analysis. *Frontiers in Communication*, **3**, 36. https://doi.org/10.3389/fcomm.2018.00036.

Sebastian, C. L., **Fontaine, N. M.**, **Bird, G.**, **Blakemore, S. J.**, **Brito, S. A.**, **McCrory, E. J.**, & **Viding, E.** (2012). Neural processing associated with cognitive and affective theory of mind in adolescents and adults. *Social Cognitive and Affective Neuroscience*, **7**(1), 53–63. https://doi.org/10.1093/scan/nsr023.

Shamay-Tsoory, S. G., & **Aharon-Peretz, J.** (2007). Dissociable prefrontal networks for cognitive and affective theory of mind: A lesion study. *Neuropsychologia*, **45**(13), 3054–3067. https://doi.org/10.1016/j.neuropsychologia.2007.05.021.

Shamay-Tsoory, S. G., **Shur, S.**, **Barcai-Goodman, L.**, **Medlovich, S.**, **Harari, H.**, & **Levkovitz, Y.** (2007). Dissociation of cognitive from affective components of theory of mind in schizophrenia. *Psychiatry Research, 149*(1-3), 11–23. https://doi.org/10.1016/j.psychres.2005.10.018

Sotomayor-Enríquez, K., **Gweon, H.**, **Saxe, R.**, & **Richardson, H.** (2023). Open dataset of theory of mind reasoning in early to middle childhood. *PsyArXiv*, **52**, 109905. https://doi.org/10.31234/osf.io/gczp9.

Tiv, M., **O'Regan, E.**, & **Titone, D.** (2021). In a bilingual state of mind: Investigating the continuous relationship between bilingual language experience and mentalizing. *Bilingualism: Language and Cognition*, **24**(5), 918–931. https://doi.org/10.1017/S1366728921000225

van der Meer, L., **Groenewold, N. A.**, **Nolen, W. A.**, **Pijnenborg, M.**, & **Aleman, A.** (2011). Inhibit yourself and understand the other: Neural basis of distinct processes underlying theory of mind. *NeuroImage*, **56**, 2364–2374. https://doi.org/10.1016/j.neuroimage.2011.03.053.

Verhagen, J., **Mulder, H.**, & **Leseman, P. P. M.** (2017). Effects of home language environment on inhibitory control in bilingual three-year-old children. *Bilingualism: Language and Cognition*, **20**(1), 114–127. https://doi.org/10.1017/s1366728915000590

Wechsler, D. (1974). *Wechsler intelligence scale for children–revised*. Psychological Corporation.

Wellman, H. M. (2018). Theory of mind: The state of the art. *European Journal of Developmental Psychology*, **15**(6), 728–755. https://doi.org/10.1080/17405629.2018.1435413.

Wellman, H. M., **Cross, D.**, & **Watson, J.** (2001). Meta-analysis of theory-of-mind development: The truth about false belief. *Child Development*, **72**(3), 655–684. https://doi.org/10.1111/1467-8624.00304.

Wellman, H. M., & **Liu, D.** (2004). Scaling of theory-of-mind tasks. *Child Development*, **75**(2), 523–541. https://doi.org/10.1111/j.1467-8624.2004.00691.x.

Wimmer, H., & **Perner, J.** (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in children's understanding of deception. *Cognition*, **13**, 103–128. https://doi.org/10.1016/0010-0277(83)90004-5.

Yott, J., & **Poulin-Dubois, D.** (2016). Are infants' theory of mind abilities well integrated? Implicit understanding of intentions, desires, and beliefs. *Journal of Cognition and Development*, **17**(5), 642–658. https://doi.org/10.1080/15248372.2015.1086771.

Yu, C.-L., **Kovelman, I.**, & **Wellman, H. M.** (2021). The development of explicit theory of mind across childhood: New insights from bilingualism. *Child Development Perspectives*, **15**(3), 154–159. https://doi.org/10.1111/cdep.12412.

https://doi.org/10.1017/S1366728926101229 Published online by Cambridge University Press