Bilingualism: Language and Cognition

cambridge.org/bil

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

Marta Białecka[1,*], Zofia Wodniecka[2,*], Karolina Muszyńska[3], Marta Szpak[2] and Ewa Haman[3]

[1]Nicolaus Copernicus University, Torun, Poland; [2]Jagiellonian University, Krakow, Poland and [3]University of Warsaw, Warsaw, Poland

Research Article

*joint first authorship

Cite this article: Białecka, M., Wodniecka, Z., Muszyńska, K., Szpak, M., & Haman, E. (2023). 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*, 1–19. https://doi.org/10.1017/S1366728923000652

Received: 3 May 2022
Revised: 28 June 2023
Accepted: 6 August 2023

Keywords:
bilingualism; theory of mind; children

corresponding author:
Marta Białecka;
Email: marta.bialecka@umk.pl

Abstract

Previous studies that contrasting bilinguals with monolinguals on Theory of Mind (ToM) have shown mixed results. We present a relatively large ($N = 102$) study comparing Polish–English sequential bilinguals living in the UK with Polish monolinguals living in Poland. Going beyond a simple group comparison, we explored the role of language proficiency and input in ToM abilities. A battery of eight tasks was used to measure ToM, and the groups were matched on age, gender, SES, IQ and L1 word comprehension. Although bilinguals did not differ from monolinguals in accuracy in ToM tasks, they demonstrated better reasoning abilities when providing justification for ToM responses. ToM accuracy scores were best predicted by L1 proficiency, but the justification scores were best predicted by both L1 and L2 proficiency. The findings suggest that the nuances of bilingual experience provide an important scaffolding context for ToM development.

Introduction

Theory of Mind (ToM) is usually defined as the ability to attribute mental states to others in order to predict or explain their behavior (for a review, see Sabbagh & Bowman, 2018). The large body of ToM research on monolingual children gathered over the last 40 years has established various cognitive and social predictors of ToM development at preschool age, e.g., Executive Functions or social abilities (see for review: Hughes & Devine, 2019). Also, skills related to language have been listed as important predictors of ToM, e.g., general language abilities, receptive vocabulary, or understanding of complement clauses (for a review, see Astington & Baird, 2005; de Villiers & de Villiers, 2014; Milligan et al., 2007; for a meta-analysis, see Tompkins et al., 2019). Despite a clear connection between language abilities and ToM, the potential impact of bilingualism on ToM has been addressed in relatively few studies; recent reviews report only 16 (Schroeder, 2018) or 24 (Yu et al., 2021) such studies.

Given the importance of linguistic abilities for ToM development in monolinguals, it seems reasonable to expect that bilingualism (i.e., the knowledge and the experience of living within a two-language environment) should exert some impact on how ToM develops. The seminal study of Peggy Goetz (2003) brought the first empirical support for an advantage of bilingual children in ToM. Since then, at least 13 studies with children below the age of seven years but above the age of four have explored this idea further. Several of these studies provided evidence in favor of the idea that bilingualism might indeed lead to earlier development of ToM, although negative evidence is also available (for a systematic review, see Rubio-Fernández, 2017; for a meta-analysis, see Schroeder, 2018; Yu et al., 2021). However, still little is known about the underlying mechanism(s) and the predictors of ToM ability in bilinguals. Especially lacking is an understanding of the role that language skills as well as L1 and L2 input play in ToM development.

It is undisputed that growing up with two languages naturally establishes a different linguistic environment and different computational demands than growing up with just one language. In other words, not only language skills (e.g., proficiency in two languages) but also experience with more than one language (e.g., the quantity and quality of input) may impact ToM abilities. Therefore, the main goal of this paper is to address the important gaps in the literature and explore how “language factors” – which are considered critical in ToM

CAMBRIDGE
UNIVERSITY PRESS

This article has earned badges for transparent research practices: Open Data. For details see the Data Availability Statement.

development in monolingual children (see Hughes & Devine, 2019 for a review; Milligan et al., 2007 for a meta-analysis) – contribute to ToM reasoning in bilinguals aged 4–6.

Theory of Mind in monolinguals: critical predictors and pitfalls in assessment

ToM, or 'mindreading ability', refers to the attribution of mental states, i.e., beliefs, thoughts, feelings, desires, emotions, or intentions, to others in order to predict or explain their behavior. This ability has been widely investigated for almost 40 years in monolinguals, both adults and children (see, e.g., Apperly, 2011; Hughes & Devine, 2019; Wellman, 2014 for a review). In younger children, before and just after their fourth birthday, a standard first-order false-belief task (Wimmer & Perner, 1983) has typically been used. Although the use of only one type of task to measure a complex ability may raise a reliability of measurement issue (Hughes et al., 2000), it is only recently that sets of tasks or scales have been developed for toddlers and young children (e.g., Białecka-Pikul et al., 2018; Wellman et al., 2011). For older children (aged 4–6) who pass the first-order false-belief task, various other tasks have been developed, each of which probably taps slightly different ToM abilities.

The false-belief tasks commonly used with four-year-olds and younger children are the unexpected transfer task (Wellman et al., 2001) and the unexpected content task (Perner et al., 1987). These are first-order belief tasks as they require the tested child to consider the beliefs of a story character. Usually, children successfully pass these tasks at age four (Wellman et al., 2001). Importantly, both of these tasks are "pass-fail" tasks, meaning there is a 50% likelihood that a particular child passes the task by chance, which creates problems with interpreting results. One way to overcome this challenge is to supplement the main test questions (e.g., Where will Max look for the chocolate?) with a justification question (e.g., Why will he look there?). However, previous research has rarely used this strategy (for an exception, see, e.g., Białecka-Pikul et al., 2018). Such a direct question may be very helpful as it provides information on whether a child can explicitly refer to the mental states of others (e.g., that the story character DOES NOT KNOW that Mum moved the chocolate, or he THINKS the chocolate is there) and explain a character's reasoning process. Thus, by requiring children to justify their answers, we gain better insight into a child's thinking process and conceptual development (Lombrozo, 2006); therefore, we can measure ToM reasoning more accurately. Naturally, "why" questions put higher linguistic demands on a child than "what" or "where" questions (de Villiers, 1991).

Another solution for improving measurement of ToM in older children is to use a series of tasks of varied difficulty. More difficult ToM tasks tap into recursive thinking, i.e., thinking about thinking abilities (Miller, 2012). For example, in the second-order false-belief task (Perner & Wimmer, 1985; Tager-Flusberg & Sullivan, 1994) that is used with five- and six-year-olds, the child is required to consider a character's belief about another character's belief (e.g., Mum thinks that Max thinks that…). Other tasks for older children measure interpretative abilities (Chandler & Lalonde, 1996), such as understanding of interpretation (Lalonde & Chandler, 2002), understanding of ambiguity (Carpendale & Chandler, 1996), deception (Talwar et al., 2007), or understanding of somebody's surprise (Hadwin & Perner, 1991). All these tasks are definitely more complex than standard first-order stories, and they impose greater linguistic demands. As such, they provide a more sensitive assessment of ToM development in children older than four years.

The importance of language skills for ToM performance has been demonstrated in longitudinal studies as well as in research on atypical populations (e.g., Mazza et al., 2017) or intervention studies (e.g., Lohmann & Tomasello, 2003). Reviewing all the studies that have used this methodological perspective is beyond the scope of the paper; however, in short, a meta-analysis by Milligan et al. (2007) found that language abilities (e.g., semantics or syntax) explained a large portion of variance in ToM (effect size, $r = .43$). In all the longitudinal studies included in this meta-analysis, the relation between early language ability and later ToM was stronger than the opposite (i.e., early ToM and later language ability), which suggests that language provides a foundation for ToM development – not the other way round (Hughes & Devine, 2019). In sum, in monolingual children, the better the language abilities, the more enhanced the ToM development, at least in 4 year-olds, who have not started systematic and formal language education.

ToM abilities in monolinguals have also been shown to be impacted by factors such as 1) age (e.g., Wellman et al., 2001; a critical change in false-belief understanding occurs in four-year-olds); 2) gender (e.g., Walker, 2005, girls outperformed boys); 3) socio-economic status (e.g., Devine & Hughes, 2018; the higher the status, the better the ToM, $r = .18$); and 4) 'executive function' (see the meta-analysis by Devine & Hughes, 2014 – the more developed the EF, the higher the ToM, $r = .38$)[1]. All these factors should be taken into account or at least controlled for in research on ToM reasoning in children in early and middle childhood.

Theory of Mind in bilinguals: state of the art

As has been indicated, there are grounds to suggest that bilingual children may develop ToM abilities earlier than their monolingual peers. As stated initially by Goetz (2003) and more recently by Schroeder (2018) and Yu et al. (2021), a bilingual advantage in ToM should be expected for at least three different reasons, all of which are grounded in research comparing bilinguals with monolinguals. First, some studies suggest that bilinguals demonstrate greater meta-linguistic awareness than monolinguals, possibly because representations stored in two languages strengthen their general meta-representational skills (e.g., Doherty, 2000). In other words, using two languages to communicate with others can help (or even be essential for) metalinguistic and metacognitive abilities, including reasoning about our own and other people's thinking processes. Second, bilinguals show greater socio-linguistic or pragmatic abilities (by being able to adjust their language to others even at the age of two – e.g., Genesee et al., 1996), as well as an enhanced ability to follow the perspective of the interlocutor while communicating. Thus, these communicative abilities may in turn enhance thinking about the content of other people's minds. Finally, bilingual children often demonstrate more enhanced cognitive control abilities than monolinguals (e.g., Bialystok & Craik, 2010), which in themselves may provide a necessary or just an important factor supporting the bilingual advantage in ToM.

Schroeder (2018) conducted a meta-analysis of 16 studies that compared ToM performance between bilingual and monolingual children ($N = 1283$) and revealed a small bilingual advantage in ToM ability (Cohen's $d = 0.22$). The effect reached medium size (Cohen's $d = 0.58$) when the transformed ToM scores were

https://doi.org/10.1017/S1366728923000652 Published online by Cambridge University Press

statistically adjusted for bilingual vs monolingual differences in language proficiency. As argued by Schroeder, the results provide support for a beneficial effect on ToM reasoning of acquiring two languages. However, the studies included in Schroeder's meta-analysis differed in various factors that were not accounted for in the analysis. These included 1) the selection and matching of bilingual and monolingual samples; 2) the type of children's exposure to the second language (e.g., simultaneous vs. sequential); 3) the age range of the tested children; 4) the particular tasks used to measure ToM; 5) the L1 and L2 skills of bilinguals (if both were tested); and 6) the language in which these children were tested. It is therefore unclear to what extent the large heterogeneity across the studies contributed to the relative weakness of the observed main effect in ToM abilities and why, as suggested by Schroeder (2018), adjusting ToM scores "for bilingual-monolingual differences in language proficiency" (p.8) enhanced the strength of the effect.

More recently, Yu et al. (2021) reviewed 24 studies investigating the relation between ToM and bilingualism. Echoing the conclusion of Schroeder (2018), Yu et al. state that the bilingual advantage for ToM development appears modest. These authors also suggest that meta- or socio-linguistic accounts provide a more plausible albeit less studied explanation of the ToM advantage than accounts that assume the critical role of executive functions. In our opinion, the conclusion formulated by Yu et al. is premature because there are no longitudinal or experimental studies that have directly investigated how experience with two languages impacts development of ToM. Importantly, all three explanations can be in fact complementary in many ways. For example, socio-linguistic skills can impact ToM directly or via meta-linguistic skills. Alternatively, EF can impact ToM directly or be a mediator between socio-linguistic factors and ToM.

Below, we provide a more in-depth qualitative analysis of the previous research, focusing on the role of the linguistic abilities and language exposure that may have played a crucial role in the outcomes of that research. We constrained our analysis to studies with children older than four years of age but younger than seven years of age. We assume that these children can pass first-order FBU tasks and start passing more sophisticated tasks (e.g., second-order FBU). Moreover, at around the age of seven, children start using "language for learning" and thus develop metalinguistic skills which probably impact ToM abilities[2]. Additionally, as mentioned above, ToM abilities change quite substantially between the ages of four and seven (see Astington & Hughes, 2013).

Studies on ToM in bilinguals – similarities and differences in methodology

Our review includes thirteen published studies[3] out of which eight found a bilingual advantage in ToM and five did not. Our aim was to focus on factors which were different across the studies and consider whether these differences might have biased the outcomes. We identified four such factors: participant age and the age range, the sample-matching strategy, the type and number of ToM tasks used, and language proficiency. Below, we summarize the outcomes of our review, while considering each of the factors (see Table 1 for details of the review).

Participant age and the age ranges across the studies

The tested samples were relatively small (the smallest $N = 14$) to moderate (the biggest $N = 98$). The total age range of the tested children was 2 years 1 month (notation: 2;1) to 6;10; the assessed children's mean age in nine studies was 4;4. Notably, these wide age ranges make a between-studies comparison of ToM and the language abilities of the tested children quite problematic.

Matching strategy

In all the studies, the compared groups were matched on age. Five studies additionally matched the groups on gender, while eight studies gave no information on matching by gender. In seven studies, parental SES or the level of education were comparable between groups; in three studies, the parents of monolingual children had higher SES than the parents of bilinguals; and in three studies, no information regarding SES was provided. Overall, it is clear that the compared samples of bilinguals and monolinguals were not fully matched on the sociodemographic variables that may impact ToM (Hughes et al., 2005).

Type and number of ToM tasks used

In eleven of the thirteen studies included in our analysis, one to six standard ToM tasks (false-belief understanding tasks: deceptive box task or unexpected transfer task, appearance-reality task) were used. In two studies, the tasks measured not ToM per se but social communication abilities (Fan et al., 2015) or cognitive perspective taking (Han & Lee, 2013). The differences in the executive as well as linguistic demands of the different tasks used to measure ToM make it difficult to provide any general conclusion regarding how such a complex ability as ToM relates to another complex ability and/or context of learning ToM as language. Moreover, when measuring ToM with the use of one or two "pass vs. not passed" tasks, the reliability of such measurement is disputable.

Language as a factor in study design or in analysis of results

The language of testing seems to be rather critical for the accurate assessment of ToM abilities not only because – as we argued earlier – ToM ability, even in monolinguals, is impacted by language ability, but also because bilinguals' language skills in each language are typically lower than those of age-matched monolinguals (Bialystok et al., 2010; Bonifacci et al., 2017; Haman et al., 2017; Hoff et al., 2014). Our review revealed that only three of the thirteen reviewed studies tested bilingual participants for ToM abilities in both languages which they knew (see Table 1). If, we recognize that language not only serves to reveal ToM, but also allows ToM's development (see Moses, 2001), then studying ToM in two languages of a bilingual child seems critical.

Moreover, in only three of the ten remaining studies in which bilinguals' ToM was tested in one language, bilingual children were tested in their dominant (as objectively tested) or preferred (as pointed by parents) language; in five they were tested in their L2 (language of formal education); and in the other two studies they were tested in their L1 (home language). Again, even when bilinguals' ToM is tested in the dominant or preferred language, bilinguals could be expected to perform lower than their monolingual peers because they typically have smaller language skills in each of their languages compared to monolinguals (e.g., Haman et al., 2017; Łuniewska et al., 2022). Importantly, it is also difficult to speculate if home language – and, in general, home environment – is more or more or less important for ToM development than the language input and skills acquired in the education system (this depends of the characteristics of daycares, time spent there and quality of interactions provided by such institutions). Importantly, in all of the reviewed studies, the precise

Table 1. Studies comparing ToM in bilingual and monolingual children. Part A) list of studies reporting a bilingual advantage; Part B) list of studies reporting a bilingual advantage only after controlling for language proficiency; Part C) list of studies reporting no difference between the groups

AUTHOR *(by year of publication)*	TESTED GROUPS *(languages and context of use (if available),sample size and age range of participants)*	CRITERIA of RECRUITMENT or INCLUSION IN THE ANALYSIS	VARIABLES CONTROLLED FOR	ToM TASKS	LANGUAGE OF TESTING	LANGUAGE ABILITIES: MEASURES USED AND MAIN RESULTS	ToM RESULTS *(accuracy)*
A) Studies reporting bilingual advantage in ToM							

Goetz (2003)	Bilinguals:
	– Mandarin-English,
	– *N* = 40,
	– range 3;2–4;1[a],
	– Mandarin at home,
	– English at daycare.
	Monolinguals:
	– English,
	– *N* = 32,
	– range 3;2–4;10.
	Monolinguals:
	– Mandarin Chinese, – *N* = 32,
	– range 3;2–4;10.
		All groups recruited and tested in daycare centers:
		– in the US (home language was Mandarin or American English),
		– in Beijing (home language was Mandarin).
		Inclusion criteria for bilinguals:
		PPVT [b] score (Dunn et al., 1997) was higher than the lowest score obtained by a monolingual.
			Half of the group were 3-year-olds and half of the group were 4-year-olds; comparable distribution of males and females in both groups;
			groups matched according to the parental level of education (every child tested had at least one college-educated parent).
				Two sessions and 4 ToM tasks:
				– 2 Appearance-reality tasks;
				– 2 unexpected transfer false-belief tasks;
				2 Level 2 visual perspective-taking tasks.
					L1 and L2 for bilinguals (two sessions).
					L1 (native language) during both sessions for monolinguals.
						Vocabulary comprehension via PPVT or its translated (non-standardized) Chinese version.
						No direct information about significant differences between groups; only standard scores and standard deviations were presented; comparable results in L1 and L2 in bilinguals.
							BI > MONO overall and when language was controlled for.

Kovács (2009)	Bilinguals: – Romanian-Hungarian, – *N* = 32, – *M* = 3;4, – range 2;1–3;7. Monolinguals: – Romanian, – *N* = 32, – *M* = 3;4, – range 2;1–3;7.	Recruited from preschools (both languages spoken); in the bilingual group, the mother tongue was Romanian and parents spoke to their children in two different languages (daily exposure); equal distribution of males and females in both groups; participants were from middle- and upper middle-class families.	Groups were matched for socioeconomic status and scores on intelligence tests.	2 ToM tasks: – standard unexpected transfer false-belief task; – modified unexpected transfer false-belief task	L2 (Hungarian).	Comparable results in vocabulary scales of Wechsler Intelligence Scale.	BI > MONO in both tasks.
Farhadian et al. (2010)	Bilinguals: – Kurdish-Persian, – *N* = 98, – *M* = 4;8, – *SD* = 6 months. Monolinguals: – Persian, – *N* = 65, – *M* = 4;4, – *SD* = 5 months.	Recruited from kindergartens in the capital city of Kurdistan; unequal distribution of gender: 54 boys and 44 girls in the bilingual group and 45 boys and 20 girls in the monolingual	Stratified simple random sampling method was used to select tested children from both groups.	3 ToM tasks: – 2 standard unexpected transfer false-belief tasks; – 1 standard unexpected content false-belief task.	L2 (Persian).	No direct comparison of language skills between the monolingual and the bilingual group was provided but the means in the Persian version of the McCarthy Scales of Children Abilities (McCarthy, 1972) are presented (M_{bi}= 63.39	BI > MONO in sum of points of all three tasks; hierarchical regression showed that linguistic status (BI vs MONO) predicted ToM when age and

		group; children were from different socioeconomic classes.				points and $M_{mono} = 55.29$ points) and suggested better language skills in bilinguals. ToM scores correlated significantly with language skills (.32 in the bilingual group; .58 in the monolingual group).	language skills were controlled for.
Fan et al. (2015)	Two groups of bilinguals (exposure and users): – English-Other, – each *N* = 24, – each *M* = 5;4, – range 4;0–6;10. Monolinguals: – English, – *N* = 24, – *M* = 5;5, – range 4;0–6;8.	Maternal education and family income did not differ across the language groups; all children lived in the greater Chicago area, were recruited from a database for psychology research, and were tested in the Lab.	No matching procedure was used; no information about gender within groups.	Social communication task (The Director Task).	L1 (English).	Vocabulary scores on the PPVT-4 (Dunn & Dunn, 2007) were not significantly different across the groups.	BI > MONO Overall and when language was controlled for.
B) Studies reporting a bilingual advantage only when language proficiency was controlled for							
Bialystok & Senman (2004, study 2)	Bilinguals: – Other-English, – *N* = 43, – *M* = 4;10, – *SD* = 10 months. Monolinguals: – English – *N* = 52, – *M* = 4;9, – *SD* = 9 months.	All groups were recruited and tested in schools; comparable distribution of males and females in the whole sample.	Half of the group were 4-year-olds and half of the group were 5-year-olds; socioeconomic status of bilinguals was lower than that of monolinguals	4 Appearance-reality tasks	L2 (English).	PPVT-R (Dunn & Dunn, 1981) scores were used as a covariate in the analysis. Monolinguals scored higher than bilinguals	BI vs MONO – no difference but BI > MONO only when language was controlled for.
Nguyen & Astington (2014)	Bilinguals: – English-French, – *N* = 24, – range 3;0–5;0. Monolinguals: – English, – *N* = 24, –range 3;0–5;0. Monolinguals: – French, – *N* = 24, – range 3;0–5;0.	Inclusion criteria: – bilinguals were exposed to both English and French before 8 months, – bilinguals exposed to each language for a minimum of 30% of the time, – monolinguals exposed only to their native language from birth, – monolinguals exposed to English/ French 90% of the time. All children had to have a minimum score on a verbal ability test (PPVT).	No difference between groups in socio-economic status and parental education. No information about gender within groups.	2 ToM tasks: – Unexpected transfer task; – Unexpected content task.	L1 and L2 for bilinguals (two sessions). L1 (native language) during both sessions for monolinguals.	Bilinguals had lower scores than monolinguals in PPVT in both languages (English/French).	BI vs MONO English – no difference, BI vs MONO French – no difference, but BI > MONO English only when language and age were controlled for.

(*Continued*)

Table 1. (*Continued.*)

AUTHOR *(by year of publication)*	TESTED GROUPS *(languages and context of use (if available),sample size and age range of participants)*	CRITERIA of RECRUITMENT or INCLUSION IN THE ANALYSIS	VARIABLES CONTROLLED FOR	ToM TASKS	LANGUAGE OF TESTING	LANGUAGE ABILITIES: MEASURES USED AND MAIN RESULTS	ToM RESULTS *(accuracy)*

Diaz & Farrar (2018a time1)	Bilinguals:
	– Spanish–English,
	– *N* = 40,
	– *M* = 4;1,
	– *SD* = 7 months.
	Monolinguals:
	– English,
	– *N* = 38,
	– *M* = 4;0,
	– *SD* = 7 months.
		Children recruited from preschools in predominantly bilingual and predominantly monolingual communities in the US.
		Inclusion criteria:
		– bilingual children identified by their parents as fluent in both Spanish and English and regularly interacting with speakers of both languages;
		the majority of the bilinguals (77.5%) had been exposed to both languages since birth.
			No differences between bilinguals and monolinguals in socioeconomic status (SES), as reflected in maternal education, occupation, and income;
			no information about gender within groups.
				3 ToM tasks were used:
				– Unexpected transfer task;
				– Unexpected content task;
				– Appearance-reality tasks
					Bilinguals tested in dominant language, as indicated by parents (37% of the group in Spanish, L1).	Receptive One Word Picture Vocabulary Test, (Gardner, 1985) and Expressive One Word Picture Vocabulary Test (Bronwell, 2000) were used in both languages with bilinguals.
						Monolingual children outperformed bilinguals in both receptive and expressive vocabulary.
							BI vs MONO – no difference but BI > MONO only when language was controlled for.

Diaz & Farrar (2018b)	Bilinguals: – English-Spanish, – *N* = 32, – *M* = 4;2, – *SD* = 7 months. Monolinguals: – English, – *N* = 33, – *M* = 4;2, – *SD* = 6 months.	Bilinguals fluent in Spanish and English and regularly interacting with speakers of both languages; the majority of the bilinguals (61%) had been exposed to both languages since birth, and all had been exposed to their non-dominant language for at least one year.	Mothers of monolingual children had on average higher education level than mothers of bilingual children; there were no significant age and gender distribution differences between the groups.	4 ToM tasks: – 2 Appearance -reality tasks; – 2 standard false-belief tasks (Unexpected transfer task and Unexpected content task).	Dominant language (L1) for bilinguals (as reported by parents, teachers, and tested by researchers). Native language for English monolinguals.	Tested with Clinical Evaluation of Language Fundamentals (CELF; Wiig, Secord, & Semel, 2004); receptive vocabulary was tested with Gardner's (2000) test and with the Comprehension for complementation task (de Villiers & Pyers, 2002) in their dominant language; monolinguals outperformed bilinguals in all these tasks.	BI vs MONO – no difference but BI > MONO only when language was controlled for.
C) Studies reporting no difference in ToM between the groups							
Han & Lee (2013)	Bilinguals: – Korean-English, – *N* = 73, – range 3;2–6;7. Monolinguals: – Korean, – *N* = 60, – range 4;4–6;10.	All children lived in the Seoul metropolitan area and Busan, South Korea; recruited from kindergartens and schools; unequal distribution of gender: 30 boys	Half the children in each language group were 4-year-olds and half were 5-year-olds; parents of bilinguals were Korean.	2 tasks (Kurdek & Rodgon, 1975): – Cognitive perspective taking task; – Affective perspective taking task.	In the bilingual group, the tasks were conducted in their preferred language (Korean or English).	All children were tested with the Korean Picture Vocabulary Test (Kim et al., 1995), a test corresponding to PPVT 3rd edition (Dunn & Dunn, 1997) and English PPVT; used to screen out	In the cognitive perspective taking task, BI vs MONO – no difference. In the affective perspective taking task, BI > MONO.

		and 43 girls in the bilingual group; 32 boys and 28 girls in the monolingual group; all middle and upper class SES; raw data suggest comparable SES.				monolinguals for bilingualism; in bilinguals, fluency in both languages was balanced between their languages and their L2 was age-appropriate; no information about differences between groups and no raw data were provided.	
Kyuchukov & de Villiers (2009, study 2)	Bilinguals: – Romani-Bulgarian, – *N* = 60, – range 3;8–5;0. Monolinguals: – Bulgarian, – *N* = 60, – range 3;8–5;0.	Tested in kindergartens; no direct information about SES of the groups but Romani bilinguals may present lower SES; no information about gender.	Similar distribution of age in each group.	2 ToM tasks: – unexpected transfer false-belief task; – unexpected content false-belief task.	L1 (Romani) and L2 (Bulgarian).	No information on whether language abilities were tested.	BI vs MONO no difference when BI were tested in L1 and Bi > MONO when BI were tested in L2 but it was probably the second test.
Pearson (2013; study 4)	Bilinguals: – English-Spanish. – *N* = 28, – *M* = 3;10, – *SD* = 9 months. Monolinguals: – English, – *N* = 40, – *M* = 3;10, – *SD* = 11 months.	Tested at home or in preschool; comparable gender distribution: – 13 boys and15 girls in the bilingual group, – 20 boys and 20 girls in the monolingual group; no information about SES.	No matching procedure was used.	2 ToM tasks: – standard unexpected transfer false-belief task; – modified (with elements of deception) unexpected transfer false-belief task.	L1 (English).	Verbal mental age, as assessed by PPVT of monolinguals, was higher than that of bilinguals.	BI vs MONO no difference.

Gordon (2016)	Bilinguals:
	– Spanish–English,
	– *N* = 26,
	– *M* = 4;7,
	– range 3;0–6;2.
	Monolinguals:
	– English,
	– *N* = 26,
	– *M* = 4;6,
	– range 3;0–6;5.
		Children recruited through local schools, the University of Minnesota's Student Housing, and the Institute of Child Development Participant Pool;
		comparable distribution of gender:
		– 12 boys and 14 girls in the bilingual group,
		– the same among the monolinguals.
			Parents of participants were asked to identify children who were proficient in both English and Spanish;
			each bilingual child was matched as close in age as possible with a monolingual child (from a larger pool of 39 children) with the same gender.
				Wellman and Liu's scale (2004) or analog versions of seven ToM tasks were used:
				– diverse desires;
				– diverse beliefs;
				– knowledge access;
				– contents false-belief;
				– explicit false-belief;
				– belief-emotion;
				– real-apparent emotion.
					L2 (English).	Children were tested with PPVT (or Spanish version of the test if they chose so).
						Higher vocabulary scores in monolinguals in comparison to bilinguals.
							BI vs MONO no difference.
							Proficiency in English was an important predictor of ToM in the MONO but not in the BI group.

Dahlgren et al. (2017)	Bilinguals: – Slavic language (e.g., Serbo-Croatian)-Swedish, – *N* = 14, – *M* = 4;6, – range 3;3–5;4. Monolinguals:	Bilinguals selected from a small town with mainly one industry, consequently mostly blue-collar workers, with a lot of	Children matched on age.	Five ToM tasks were used: 1 standard unexpected transfer false-belief task and 4 tasks suitable for younger children:	L2 (Swedish).	Monolinguals outperformed bilinguals in vocabulary tested with Swedish version of PPVT.	BI vs MONO also no difference when language skills were controlled for.

(*Continued*)

Table 1. (*Continued.*)

AUTHOR (*by year of publication*)	TESTED GROUPS (*languages and context of use (if available),sample size and age range of participants*)	CRITERIA of RECRUITMENT or INCLUSION IN THE ANALYSIS	VARIABLES CONTROLLED FOR	ToM TASKS	LANGUAGE OF TESTING	LANGUAGE ABILITIES: MEASURES USED AND MAIN RESULTS	ToM RESULTS (*accuracy*)

	- Swedish
	- *N* = 14,
	- *M* = 4;4,
	- range 2;11–5;5.
		immigrants from former Yugoslavia;
		no information about SES of monolinguals;
		all children were typically developing, as tested with Raven Matrices or equivalent for younger kids;
		no information about gender within groups.
				- Kiki and the cat (Lewis, 1994);
				- Thought picture (Woolfe et al., 2002);
				- Hide the sticker (Call & Tomasello, 1999);
				- Hide the fruit (Vinden, 1999).			

Note: [a] Age of children is denoted as mean (*M*) in number of years and months (e.g., 4;4 means 4 years 4 months) and as a range in the same notation. If available, standard deviation (*SD*) is also presented; [b] PPVT – Peabody Picture Vocabulary Test Third Edition

characteristics of children's language exposure (i.e., the quantity and quality of input) were very differently measured, thus it is difficult to compare their results.

Additionally, information about the quantity and quality of input in first and second languages was provided for only four of the 13 studies (see column 2 and 3 of Table 1). In one of the studies, the daily exposure was defined only by stating that "(a) parents of different mother tongues who each address the child in their native language; and (b) daily exposure to both languages" (Kovács, 2009, p. 50). However, in another study (Nguyen & Astington, 2014), more detailed information was provided (i.e., bilinguals were exposed to both English and French before 8 months for a minimum of 30% of the time, and monolinguals were exposed only to their native language from birth for 90% of the time). The lack of information about the language input made the control of this variable impossible. In consequence, the impact of language input on ToM has not been established.

To account for the potentially lower language skills in bilinguals, three of the thirteen reviewed studies matched the tested groups by language skills; eight studies controlled for language skills when comparing ToM in monolinguals and bilinguals. In the two of the reviewed studies, language skills were not tested at all; in one study, language proficiency was tested but the results were not provided by the authors. Overall, previous research on ToM only rarely fully controlled for proficiency and exposure in both languages when testing bilingual children, and the study designs typically did not allow the role of language proficiency and language input in the ToM abilities of bilinguals to be directly investigated.

Is the 'bilingual advantage' in ToM real or not? The role of language proficiency and exposure

The exact impact of language experience and proficiency on ToM in bilinguals is still unknown. Interestingly, some authors have put forward the idea that it is the mere fact of living in a bilingual language environment rather than the length and/or intensity of this experience that plays a role. For example, Yow and Markman (2015) argued that it is bilingual children's practice in understanding other people's linguistic perspectives which may boost their ToM reasoning, regardless of their language proficiency. Also, Fan et al. (2015) found that even "some regular but limited exposure" (p. 1091) to the second language confers an advantage in perspective-taking ability to the same degree as everyday contact with two languages. Below, we briefly summarize the types of outcomes of previous studies (no bilingual advantage, bilingual advantage, and advantage contingent on language proficiency), with a special focus on if and how the language proficiency of the tested children was measured and accounted for in the analyses (see Table 1 for details). If information on a second language input was provided in the source articles, we included it in Table 1 (column 2 and 3).

Out of the five studies that did not find differences in ToM between bilinguals and monolinguals, in three (Dahlgren et al., 2017; Gordon, 2016; Pearson, 2013) monolinguals' language abilities outperformed those of bilinguals (two other studies gave no information on language performance). This implies that there might have been a competing effect of language proficiency in the language of the ToM testing (in two studies this was L2); if this proficiency had been controlled for, it cannot be ruled out that a bilingual advantage could have been observed in ToM skills.

Out of the eight studies that found a bilingual advantage in ToM, four reported this advantage both overall and also when language abilities were controlled for. The exact pattern of results is, however, difficult to interpret: in one study (Goetz, 2003, children tested in both languages), better language skills were observed in monolinguals. In another study (Farhadian et al., 2010; children were tested in their L2), bilinguals presented better language skills than monolinguals. In two other studies, no significant differences in language abilities between bilinguals and monolinguals were observed (Kovács, 2009; Fan et al., 2015 – children were tested in L2). Thus, overall, a ToM advantage was observed when language skills were fully controlled for or when bilinguals' L2 skills were at least not lower than those of monolinguals if L2 was the language of testing.

Importantly, in the remaining four studies, a bilingual advantage in ToM was reported *only* when the impact of language skills was statistically restrained or eliminated (see Table 1). In all these studies, monolinguals outperformed bilinguals in language skills. This implies that bilingualism not only compensates for lower skills in the language of testing that are important for ToM, but it also enhances ToM development more than language abilities per se.

Our review highlights the substantial heterogeneity across the studies, which might have contributed to the reported inconsistency of the previous research findings. We identified the following limitations of the previous studies. First, in most of this research there was a wide age range of the tested children. This might have affected the outcomes in an uncontrolled way as differences in age are intertwined with differences and changes in language abilities. Second, in many of the studies, the factors that might have impacted ToM abilities (i.e., age, gender, cognitive abilities, and parental SES) were not systematically controlled for. Third, in most of the studies, ToM was tested in only one of the bilinguals' languages, and sometimes this language was the children's weaker language. Most importantly, based on previous studies it is unclear how different aspects of bilingual language experience and language skills contribute to ToM.

Current study

Our first goal was to compare Polish–English bilingual children aged 4–6 with Polish monolinguals in Theory of Mind. Our second goal was to better understand the extent to which language proficiency and input explain ToM abilities. Based on the review of previous findings, we formulated the following hypotheses:

H1: Polish–English bilingual children aged 4–6 outperform same-aged monolinguals in Theory of Mind.
H2: In monolinguals, proficiency in their native language relates to ToM.
H3: In bilinguals, both proficiency in L1 and L2 and input in these languages relate to ToM abilities.

While testing these hypotheses, we aimed to circumvent at least some of the limitations of the previous studies. Therefore, we (1) tested a relatively large group of children in aged 4–6; (2) we carefully matched the compared groups on variables that have been found to impact both ToM and language proficiency in monolinguals (age, gender, SES, cognitive abilities as measured with IQ); (3) we tested children's ToM in their dominant language (here: Polish as the home language); (4) we obtained data for several linguistic factors that are related to language proficiency and input; 5) we probed a wide range of ToM abilities, including first- and second-order false-belief understanding as well as accuracy and justification of children's answers in eight ToM tasks.

Our participants were Polish–English migrant children aged 4–6 in the UK, and a group of Polish monolingual peers in Poland. A battery of eight tasks was used to assess ToM (Test of Reflection on Thinking; TRT, Białecka-Pikul et al., 2018), which allowed us to calculate four indices: (1) overall accuracy index, (2) overall justification index, (3) first-order false-beliefs index, and (4) second-order false-beliefs index. We considered five predictors associated with either language proficiency or language exposure in bilinguals. The first two predictors were related to language proficiency as measured via language comprehension and based on performance scores in receptive vocabulary tests in L1 and L2. The remaining three predictors related to language exposure in bilinguals and were obtained via parental reports: length of L2 (English) exposure (in months); the accumulated language input in L1 and L2, i.e., cumulative language exposure indices based on both the total time spent in Poland and in the UK and on the amount and intensity of bilinguals' exposure to their languages in both these countries (see below and see also Haman et al., 2017).

To test our first hypothesis, we conducted a series of ANCOVAs. Then, we conducted a series of regression analyses with the four outcomes of the ToM tasks (TRT) as indices of dependent variables. For hypothesis 2, we regressed ToM on L1 proficiency (after controlling for age, gender, SES, cognitive abilities as measured with IQ); for hypothesis 3, the five predictors related to language proficiency and input (described in detail below) were provided (again, after controlling for demographic and cognitive abilities). To supplement the analyses related to the first hypothesis, we then performed Bayesian analyses (see analytical strategy for details).

Method

Participants

Participants were children who took part in a large-scale project on the linguistic and cognitive development of bilingual children (Bi-SLI-PL, see Acknowledgements for details), carried out within the European COST Action IS0804. Overall, 173 Polish–English migrant children living in the UK and 311 Polish monolingual children living in Poland were tested in the project. All children were of school entrance age (age 4–6, see footnote 2). Written parental consent and children's assent were obtained for all participants. The participants were not reimbursed, but the children received "small rewards" (books, stickers, CDs with songs/nursery rhymes). The whole procedure was evaluated and accepted by the Ethics Committee at the Faculty of Psychology, University of Warsaw.

The analyses presented in the current paper are based on the biggest possible subsamples from the group of Polish–English bilinguals and Polish monolinguals (see Haman et al., 2017 and supplementary materials, Appendix 1 for the full description of how we selected the subsamples of monolinguals and bilinguals to make them as comparable as possible in reference to the control variables). In total, data from 102 children (51 bilingual and 51 monolingual) were considered for the comparison of ToM performance in bilinguals and monolinguals. The characteristics of the overall sample and the subsamples are presented in Table 2.

https://doi.org/10.1017/S1366728923000652 Published online by Cambridge University Press

The groups were identical in terms of gender distribution (31 girls in each group), and there were no differences in their age in months: $t_{(100)} = -0.57$, $p = .570$ (two-tailed). The number of years of mothers' education were comparable for the two language groups, $t_{(100)} = -0.22$, $p = .829$ (two-tailed), and most mothers had higher education (in the bilingual group – 35 mothers, i.e., 68.6%; in the monolingual group – 37 mothers, i.e., 72.5%). Moreover, there was no difference between the groups in non-verbal IQ, $t_{(100)} = 0.294$, $p = .769$ (two-tailed). As such, the bilingual and monolingual groups were comparable in terms of basic cognitive and socio-demographic characteristics.

With regards to proficiency in L1 (Polish), indicated by the percentile score obtained in OTSR (Obrazkowy Test Słownikowy – Rozumienie, i.e., The Picture Vocabulary Test – Comprehension; Haman & Fronczyk, 2012), monolinguals scored higher than bilinguals: $t_{(100)} = 1.02$, $p = .046$ (two-tailed). In terms of the overall sample, monolinguals largely outperformed bilinguals. The matching procedure made the subsamples as similar as possible, but even though it diminished the between-group difference to the verge of significance, it was impossible to obtain equal L1 performance in both groups.

Measures and procedure

Below, we present a detailed description of the main tasks we used: TRT, which is a measure of ToM; two auditory word comprehension tests and most importantly – all measures of L1 and L2 exposure in bilinguals. Note that the complete testing battery used in this study is described in more detail elsewhere (Haman et al., 2017); only a short description of all the tasks is presented in the *Procedure* section.

ToM

Test of Reflection on Thinking (TRT, Białecka-Pikul et al., 2018)

The TRT was developed for children aged 4–6 over four years old and constitutes a battery of nine tasks (one training task and eight testing tasks) in the form of illustrated stories. More specifically, the TRT's tasks or stories assess a child's understanding of appearance-reality, first-order beliefs (i.e., the unexpected transfer test and the deceptive box test), understanding of interpretation, deception, ambiguity, understanding of surprise, and second-order beliefs. Nine stories are presented by the experimenter in a set order and are aided by pictures (two to five in each story) displayed on a laptop screen (19"). Table S1 in the Supplementary Materials shows a sample story with the accompanying pictures; a detailed description of all tasks is shown in Table S2. Each story describes the actions of two protagonists (two boys or two girls because there are two gender-related versions of TRT). To measure ToM abilities after each story, the child is asked to predict the protagonist's behavior or thoughts (e.g., "where will Evan look for the book?") and also to explain the protagonist's behavior (e.g., "why will Evan be looking there?"). In other words, in TRT two kinds of questions are asked after each story, thus two indices of ToM can be calculated: overall accuracy index and overall justification index. The overall accuracy index is a sum of points scored in the questions concerning the behavior, thoughts or emotions of the protagonists (e.g., "what will she do?", "what will she think?", "how will she feel?"). For each question, a child can score one point for a correct answer and zero points for an incorrect answer, an "I don't know" answer, or no answer. The overall justification index is the sum of points scored in the "why?" questions. A child can score one point for a correct answer without clear mental references (e.g., if a child explains the protagonist's behaviors by referring to a situation or desires), and two points when clear mental references (i.e., thoughts, knowledge, beliefs) are provided. Zero points are given for a wrong answer, an "I don't know" answer, no answer, and if the answer for the accuracy question was scored zero.

Five (out of nine) stories include control questions (memory questions, e.g., "where is the book now?"); if a child failed to provide a correct answer to the memory question, he/she received 0 points for the story. The first story in TRT serves as a training story, so the child's answer is not included in the calculations of the ToM indices. Thus, a child could score a maximum of 8 points on the ToM overall accuracy index and a maximum of 16 points on ToM overall justification index. To provide a more detailed analysis of the ToM variable, the two additional indices related to the accuracy of answers in the first- and second-order false-belief tasks were calculated. The first-order false-beliefs index is a sum of the accuracy scores of two tasks: the unexpected transfer test (story 2 in TRT) and the deceptive box test (story 3 in TRT). The second-order false-beliefs index is a sum of scores for the two second-order false-belief tasks (stories 8 and 9 in TRT).

To check the reliability of the coding system for TRT, the inter-rater reliability (for two independent coders), measured on a randomly selected subsample of monolinguals ($n = 38$), was calculated and assessed as satisfactory (kappas ranged from .84 to 1.00 for tasks on both scales). The inter-rater reliability for both indices measured with alphas ($n = 254$) was also satisfactory (.62 for the ToM accuracy index and .64 for the ToM justification index). There are also data that prove the good convergent and content validity of the TRT in monolinguals (see Białecka-Pikul et al., 2018).

Language factors

Auditory word comprehension was measured in English via the British Picture Vocabulary Scale – Third Edition (BPVS, Dunn et al., 2009), and in Polish via Obrazkowy Test Słownikowy – Rozumienie, OTSR (The Picture Vocabulary Test – Comprehension; Haman & Fronczyk, 2012). Both tests (BPVS and OTSR) are published and normed on monolingual populations and were designed to assess the comprehension of nouns, verbs, and adjectives. The two tests have similar instructions: children are presented with boards of four pictures and asked to point to the picture that appropriately depicts the target word.

Three measures of language experience in the bilingual group were used: (1) length of time of L2 (English) exposure; (2) an index of cumulative language exposure to L1; and (3) an index of cumulative language exposure to L2. All three measures were based on the information from a parental questionnaire for bilingual pre-school and early-school children, i.e., a Polish adaptation of PABIQ [Questionnaire for Parents of Bilingual Children, (Tuller, 2015); Polish adaptation by Kuś et al. (2012, unpublished)]. Parental answers provided detailed information on each child's early development and language background. The length of L2 (English) exposure was calculated as the time (in months) between the age of first contact with L2 and the time of testing. The indices of cumulative language exposure to L1 and L2 were based on the total time spent in Poland and in the UK (in the child's lifetime), as well as the amount and quality of exposure to language received in each of these countries. The indices were calculated as follows (see also Haman et al., 2017).

https://doi.org/10.1017/S1366728923000652 Published online by Cambridge University Press

Table 2. The Characteristics of the Overall Sample and the Subsample

	Bilinguals		Monolinguals		Bilinguals		Monolinguals	
	Overall sample *N* = 95 (*girls* = 75)		Overall sample *N* = 268 (*girls* = 134)		Final subsample *N* = 51 (*girls* = 31)		Final subsample *N* = 51 (*girls* = 31)	
	Mean (SD)	*Range*	*Mean (SD)*	*Range*	*Mean (SD)*	*Range*	*Mean (SD)*	*Range*
Age (months)	67.4 (8.7)	52–84	67.4 (8.2)	43–87	66.8 (9.0)	52–82	67.8 (9.1)	43–84
SES: years of mother's education	16.5 (2.9)	10–24	17.4 (2.7)	11.5–25	16.6 (3.1)	10–23	16.6 (2.8)	11.5–24
Non-verbal IQ: Raven's Colored Matrices (raw score)	21.7 (5.7)	10–34	22.0 (5.1)	10–35	22.2 (5.4)	13–33	21.9 (4.7)	11–32
Age of first L2 contact (months)	12.7 (15.7)	0–60			14.6 (16.8)	0–60		
L1 (Polish) receptive vocabulary: OTSR percentile	29.4 (25.0)	1–97	59.4 (28.7)	1–100	33.9 (22.0)	8–90	42.6 (21.9)	5–85
L2 (English) receptive vocabulary: BPVS percentile	27.9 (22.4)	2–94			25.1 (19.6)	2–74		
L2 (English) length of exposure (months)	54.92 (16.48)	8–83			52.14 (17.12)	8–80		
L1 (Polish) cumulative exposure index	252.7 (75.3)	60–402			268.2 (63.8)	132–402		
L2 (English) cumulative exposure index	154.6 (76.1)	0–366			133.0 (66.4)	0–330		
ToM: TRT overall accuracy index	4.3 (2.2)	0–8	4.4 (2.0)	0–8	4.4 (2.0)	0–8	4.2 (1.9)	0–8
ToM: TRT overall justifications index	6.7 (3.2)	0–12	6.4 (3.2)	0–12	6.9 (3.2)	0–12	6.0 (3.0)	0–12
ToM: first-order false-belief index	1.33 (0.74)	0–2	1.48 (0.73)	0–2	1.41 (0.73)	0–2	1.47 (0.73)	0–2
ToM: second-order false belief index	0.69 (0.78)	0–2	0.89 (0.77)	0–2	0.69 (0.68)	0–2	0.90 (0.73)	0–2

Note: OTSR: Obrazkowy Test Słownikowy Rozumienie (The Picture Vocabulary Test – Comprehension in Polish), BPVS: British Picture Vocabulary Scale. TRT: Test of Reflection on Thinking. The overall bilingual and monolingual samples consist of children who match a profile of a typically developing child and have performed the ToM task. The subsamples consist of children who performed the ToM task in Polish, performed the non-verbal IQ task, obtained at least the 5th percentile in the Polish word comprehension test, and had the full set of background information (e.g., age, SES, and in case of bilinguals, information on language exposure).

First, we estimated the extent of each child's exposure to each language when living in the United Kingdom on the basis of the parental questionnaire: parents reported, on a 5 point Likert scale, how often their child was addressed in English and Polish in particular communicative situations (e.g., parents talking to the child, other children talking to the child). These scores were aggregated to estimate the children's exposure to Polish and to English during their stay in the United Kingdom. The maximum score for each language was 91, which would indicate that when living in a given country (e.g., Poland) the child had no contact with English. The final index reflected the time spent in Poland and in the United Kingdom in the lifetime of each child, as well as the amount of exposure the child had received in each of these countries. The index of the cumulative exposure to Polish was calculated using the following formula: (time spent in Poland) ∗ 91[4] + (time spent in the United Kingdom) ∗ (exposure to Polish while in the United Kingdom). The actual unit of measurement used to calculate the index was the child's age in days, represented as years (in decimals). The index of cumulative exposure to English was calculated as (the time spent in Poland) ∗ 0[5] + (the time spent in the United Kingdom) ∗ (the exposure to English while in the United Kingdom).

Procedure

All children were tested individually in a quiet room: the monolingual Polish children in their preschools or in their homes in Poland; the bilingual children in their schools or in their homes in the United Kingdom (for details see Haman et al., 2017). In total, each monolingual child was tested over three to four sessions, and each bilingual child was tested in both languages over five to seven testing sessions (each lasting 45–90 min). The duration of each session depended on the child's pace. The order of the tasks in the testing sessions was counterbalanced across participants. The tasks in Polish were administered by a native speaker of Polish, while the tasks in English (not included in the present report) were administered by a native speaker or a highly proficient speaker of English. Polish and English were never tested on the same day. Each child did 14 tasks in the dominant language and eight tasks in a non-dominant language (see Supplementary Materials, Appendix 2 for the list of tasks).

Here, we report only data for the tasks described above and the Raven Scale (Jaworowska & Szustrowa, 2003), performed in the dominant language – namely, Polish.

Results

The analytical strategy

The statistical analyses are presented in the following way. First, we report the preliminary analysis that compares the matched bilingual and monolingual groups on the control variables: age, socioeconomic status, non-verbal intelligence, and Polish (L1) receptive vocabulary size. These comparisons were done with the use of frequentist inference – namely, t-tests. Next, we compared the groups on the four indices provided by TRT: overall accuracy and justification indices, and first-order and second-order false-belief ToM indices. However, we included Polish (L1) receptive vocabulary size as a controlled variable. This was done using a series of ANCOVAs with language group (bilingual, monolingual) as the grouping factor, each of the TRT indices as the dependent variable, and the percentile on the Polish (L1) comprehension test as the covariate. However, frequentist inference only provides evidence against the null hypothesis and cannot provide probabilistic evidence in favor of the alternative hypothesis. We therefore also employed Bayesian inference – namely, a Bayesian ANCOVA with a Bayes Factor[6] (Hoijtink et al., 2019).

Finally, we present the results of a series of regression analyses which looked for predictors of TRT performance in the bilingual and monolingual groups. For the overall TRT accuracy and justification indices, we used a series of hierarchical regression analyses; for the first-order and second-order false-beliefs indices (which are binary variables), we used a series of logistic regression analyses.

Theory of Mind in bilinguals and monolinguals

The results indicated that bilinguals did not outperform monolinguals in the TRT overall accuracy index ($F_{(1,99)} = 1.17$, $p = .283$, $\eta^2 = .012$). The effect of the OTSR percentile covariant was statistically significant ($F_{(2,98)} = 13.09$, $p < .001$, $\eta^2 = .117$, large effect size). We also calculated the Bayes Factors for the comparison between monolinguals and bilinguals. The difference between the groups' TRT overall accuracy index provided moderate evidence in favor of the null hypothesis ($BF_{01} = 3.60$) but virtually no evidence for the alternative hypothesis (BF_{10}) = 0.28. Thus, the Bayes Factor revealed that there was no difference between groups regarding the overall TRT accuracy index.

As for the overall TRT justification index, the groups did differ in the classic hypotheses testing ($F_{(1,99)} = 5.95$, $p = .016$, $\eta^2 = .057$, medium effect size). The effect of the OTSR percentile covariant was statistically significant ($F_{(2,98)} = 17.76$, $p < .001$, $\eta^2 = .152$, large effect size). The difference provided weak evidence in favor of the null hypothesis, $BF_{01} = 1.73$. For comparison, the evidence for the alternative hypothesis was even weaker: $BF_{10} = 0.58$. Thus, in this case, the Bayesian inference could not help us to choose one hypothesis over another.

We also checked whether there were group differences in performance in the TRT first- and second-order false-beliefs indices. The results showed that bilingual children did not differ from monolinguals in the first- or second-order false-beliefs indices: $F_{(1,99)} = 0.02$, $p = .890$, $\eta^2 = .001$ and $F_{(1,99)} = 1.27$, $p = .262$, $\eta^2 = .013$, respectively. All descriptive statistics are presented in Table 2 in the *Method* Section. The effect of the OTSR percentile covariant was statistically significant in both analysis ($F_{(2,98)} =$ 7.79, $p < .010$, $\eta^2 = .073$, medium effect size and $F_{(2,98)} = 4.42$, $p < .01$, $\eta^2 = .042$, medium effect size, respectively). The Bayes Factor indicated moderate evidence for the null hypothesis: $BF_{01} = 4.44$ in the first-order false-beliefs index; a slight preference for the null hypothesis, $BF_{01} = 1.66$, in the second-order false-belief index. In essence, the Bayesian Factors indicated moderate to weak evidence for the null hypothesis, which states that the performance on first- and second-order false-belief tasks is similar in both the bilingual and monolingual groups.

The language predictors of ToM

Separate regression models were used to analyze the predictors of ToM performance for each language group, controlling for age, gender, socioeconomic status, and non-verbal IQ. For the monolingual group, only the auditory word comprehension test (OTSR) was added to the regression model. For bilingual children, a series of regression models with different language predictors were calculated: (a) L1 (Polish) and L2 (English) word comprehension (as bilinguals scored lower than monolinguals); (b) L2 length of exposure; (c) cumulative language exposure to L1; and (d) cumulative language exposure to L2.

Monolinguals: regression models for overall TRT accuracy and overall TRT justification indices

Overall TRT accuracy index

The hierarchical regression analysis revealed that sociodemographic variables accounted for 43% of the variation in the TRT overall accuracy index: $F(4,46) = 10.44$, $p < .001$. Among these variables, age, $\beta = .34$, $p = .013$, and non-verbal IQ, $\beta = .41$, $p = .003$, were significant predictors (see Table S3 in Supplementary materials). After adding the L1 word comprehension index in Step 2, the total variance explained by the model as a whole was 49%, and the model was statistically significant, $F(5,44) = 10.59$, $p < .001$. The inclusion of the L1 word comprehension index explained an additional 6% of variance in the overall TRT accuracy index, $\Delta R^2 = .06$, $F(1,45) = 6.34$, $p = .015$. In the final adjusted model, age, $\beta = .45$, $p = .001$, non-verbal IQ, $\beta = .28$, $p = .044$, and L1 word comprehension index, $\beta = .29$, $p = .015$, were significant predictors of the overall TRT accuracy index.

Overall TRT justification index

As regards the overall TRT justification index, the base model with sociodemographic variables was statistically significant, $F(4,46) = 4.68$, $p = .003$, and accounted for 23% of the variance in the overall TRT justification index. In this model, only non-verbal IQ significantly predicted the quality of bilingual children's justifications in TRT: $\beta = .41$, $p = .010$. Adding the L1 word comprehension index explained an additional 7% of variance, $\Delta R^2 = .07$, $F(1,45) = 5.44$, $p = .024$, and the final model was statistically significant, $F(4,45) = 5.19$, $p = .001$. In this model, the L1 vocabulary comprehension index, $\beta = .32$, $p = .024$, was the only significant predictor of the overall TRT justification index.

https://doi.org/10.1017/S1366728923000652 Published online by Cambridge University Press

Monolinguals: logistic regression models for the TRT first- and second-order false-beliefs indices

Two logistic regression models were used to identify the predictors of the monolinguals' performance on the TRT first- and second-order false-beliefs indices. In order to run a logistic regression model, the scores from these two indices were transformed from three-level factors (0–12– points) to two-level factors (0 points vs. > 0 points). When reporting statistically significant predictors, we report only those predictors which were significant and for which the 95% confidence interval (CI) for the odds ratio (OR) did not include 1 (if CI of OR includes 1, it means there is no association between the predictors and the outcome).

First-order false-beliefs index

The first model that was used to predict performance on the first-order false-beliefs index (i.e., whether the children scored any points at all or none) included all the sociodemographic variables: gender, age, non-verbal IQ and socio-economic status. The analysis revealed only a significant effect of non-verbal IQ ($b = 0.27$, $SE = 0.13$, OR = 1.31, $p = 0.047$). The second model included the sociodemographic variables and the L1 word comprehension index. However, the effect of language proficiency was non-significant ($b = 0.31$, $SE = 0.19$, OR = 1.36, $p = 0.107$). Model 2 showed a lower Akaike's Information Criterion (AIC) value, which suggests a more parsimonious model (AIC estimates the quality of each model). We followed the model selection criteria set out by Burnham and Anderson (2004): we calculated the difference in AIC values between each model and the model with lowest AIC. The greater the difference, the less likely it is that the model is the best approximating model among the candidates in the set.. Model 2 showed a difference in AIC values larger than 10 (ΔAIC = 14.6), which yields essentially no support for the model as being the best approximating model in the candidate set. Details of the full models are provided in Table S4.

Second-order false-beliefs index

The base model that included the sociodemographic variables was run for the second-order false-beliefs index; it revealed significant effects of gender (girls scored higher than boys, $b = 1.71$, $SE = 0.84$, OR = 5.55, $p = 0.041$) and non-verbal IQ ($b = 0.26$, $SE = 0.12$, OR = 1.29, $p = 0.030$). Model 2, extended by the L1 word comprehension index, revealed the same pattern of results: there were only two significant effects of gender ($b = 1.65$, $SE = 0.85$, OR = 5.21, $p = 0.041$) and non-verbal IQ ($b = 0.24$, $SE = 0.12$, OR = 1.27, $p = 0.030$). The effect of language proficiency was non-significant ($b = 0.01$, $SE = 0.02$, OR = 1.01, $p = 0.537$). Model 2 showed a slightly higher AIC than Model 1. The difference in AIC between Model 2 and Model 1 was smaller than 2 (ΔAIC = 1.62), which provides substantial evidence that Model 2 was the best approximating model in the candidate set. Details of the full models are provided in Table S4.

Bilinguals: regression models for the TRT overall accuracy and justification indices

The base model (Step 1 in all further regressions) that included the sociodemographic and cognitive variables explained 26%, $F(4,46) = 5.46$, $p = .001$, and 41%, $F(4,46) = 9.68$, $p < .001$, of variance in the TRT overall accuracy index and the TRT overall justification index, respectively. As regards the overall accuracy index, only age was a significant predictor, $\beta = .38$, $p = .009$; for the overall justification index, both age, $\beta = .36$, $p = .005$, and non-verbal IQ, $\beta = .42$, $p = .001$, were significant.

Overall TRT accuracy index: (a) Role of L1 and L2 word comprehension.

In the case of the TRT overall accuracy index, when the L1 word comprehension score was added in Step 2 (model 2), the total variance explained by the model as a whole was 34%, $F(5,45) = 6.25$, $p < .001$, and the change in the explained variance was statistically significant, $\Delta R^{2\ 2} = .08$, $F(1,45) = 6.83$, $p = .012$. Although model 3 (Step 3) with the L2 word comprehension score as a predictor was also statistically significant, $F(6,44) = 5.65$, $p < .001$, the change in the explained variance was non-significant, $\Delta R^2 = .02$, $F(1,45) = 1.97$, $p = .168$. Indeed, in the final adjusted model, only age and the L1 word comprehension index were significant predictors of the overall TRT accuracy index (see Table S5).

Overall TRT accuracy index: (b) Role of length of English (L2) exposure

Adding second-language experience (as measured by length of time of English (L2) exposure) to the model did not increase the explained variance in the overall TRT accuracy index, $\Delta R^2 = .02$, $F(1,45) = 1.28$, $p = .265$. However, in Step 3, adding L1 comprehension increased the explained variance in the overall TRT accuracy index, $\Delta R^2 = .09$, $F(1,44) = 7.14$, $p = .011$, and the final model was statistically significant for the accuracy index, $F(6,44) = 5.60$, $p < .001$. The overall accuracy index was predicted by age and L1 comprehension (Table S6).

Overall TRT accuracy index: (c) The role of cumulative language exposure to first and second language.

Adding L1 cumulative language exposure to the model in Step 2 did not increase the explained variance in the overall TRT accuracy index, $\Delta R^2 = .02$, $F(1,45) = 1.37$, $p = .247$, although the final model was statistically significant $F(5,45) = 4.68$, $p = .002$. Similarly, adding L2 cumulative language exposure in Step 3 did not increase the explained variance in the overall accuracy index, $\Delta R^2 = .02$, $F(1,45) = 1.28$, $p = .265$. However, adding L1 comprehension in Step 4 increased the explained variance in the overall TRT accuracy index, $\Delta R^2 = .10$, $F(1,43) = 8.41$, $p = .006$. The final model was statistically significant for the accuracy index, $F(7,43) = 5.09$, $p < .001$. Thus, the overall TRT accuracy index was predicted by age and L1 comprehension (Table S7).

Overall TRT justification index: (a) Role of L1 and L2 word comprehension.

With regards to the overall TRT justification index, adding the L1 word comprehension score in Step 2 increased the explained variance to 48%, $\Delta R^2 = .07$, $F(1,45) = 7.56$, $p = .009$, and adding the L2 word comprehension index to the model in Step 3 resulted in an additional 6% of explained variance, $\Delta R^{2\ 2} = .06$, $F(1,44) = 6.00$, $p = .018$. In Step 2, $F(5,45) = 10.36$, $p < .001$, and Step 3, $F(6,44) = 10.60$, $p < .001$, both models were statistically significant. In the final model, three variables were significant predictors of the overall TRT justification index: age, L1 word comprehension index, and L2 word comprehension index (see Table S5).

Overall TRT justification index: (b) Role of length of English (L2) exposure.

Adding second-language experience (as measured by length of time of English (L2) exposure) to the model did not increase the explained variance in the overall TRT justification index,

https://doi.org/10.1017/S1366728923000652 Published online by Cambridge University Press

$\Delta R^2 = .01$, $F(1,45) = 0.87$, $p = .356$. However, in Step 3, adding L1 comprehension increased the explained variance in the overall TRT justification index, $\Delta R^2 = .08$, $F(1,44) = 7.91$, $p = .007$. The final model was statistically significant for the justifications, $F(6,44) = 8.91$, $p < .001$. The overall justification index was predicted by age, non-verbal IQ, and L1 comprehension (Table S6).

Overall TRT justification index: (c) The role of cumulative language exposure to first and second language.

Adding L1 cumulative language exposure to the model in Step 2 did not increase the explained variance in the overall justification index, $\Delta R^2 = .01$, $F(1,45) = 1.12$, $p = .295$, although the final model was statistically significant, $F(5,45) = 7.99$, $p < .001$. Similarly, adding L2 cumulative language exposure in Step 3 did not increase the explained variance in the overall justification index, $\Delta R^2 = .01$, $F(1,44) = 1.20$, $p = .278$. However, adding L1 comprehension in Step 4 increased the explained variance in the overall TRT justification index, $\Delta R^2 = .10$, $F(1,43) = 9.87$, $p = .003$. The final model was statistically significant for the justifications, $F(7,43) = 8.51$, $p < .001$. The overall TRT justification index was predicted by age, non-verbal IQ and L1 comprehension (Table S7).

Bilinguals: logistic regression models for the first- and second-order false-beliefs indices

A series of logistic regression models were constructed to identify the predictors of the bilinguals' performance on the TRT first- and second-order false-beliefs indices. The scores from the first- and second-order false-beliefs tasks were transformed from three-level factors (0–12– points) to two-level factors (0 points, vs. points > 0).

First-order false-beliefs index

The base model for predicting performance on the first-order false-beliefs index (i.e., whether the children scored any points at all or none) included all the sociodemographic and cognitive variables, i.e., gender, age, socio-economic status, and non-verbal IQ. The analysis only revealed a significant effect of gender ($b = 2.70$, $SE = 1.26$, OR = 14.94, $p = 0.032$). In Model 2, the effect of the L1 word comprehension index was non-significant ($b = 0.02$, $SE = 0.03$, OR = 1.02, $p = 0.546$). In Model 3, the effect of the L1 word comprehension index was still non-significant ($b = 0.03$, $SE = 0.05$, OR = 1.03, $p = 0.548$), but the effect of L2 word comprehension was on the verge of significance ($b = 0.21$, $SE = 0.10$, OR = 1.23, $p = 0.046$). In Model 4, the effect of the length of English exposure was non-significant ($b = 0.02$, $SE = 0.03$, OR = 1.02, $p = 0.571$). In Model 5, the effect of L1 cumulative language exposure was non-significant ($b = -0.01$, $SE = 0.01$, OR = 0.99, $p = 0.433$). In Model 6, the effect of L1 cumulative language exposure was non-significant ($b = -0.01$, $SE = 0.01$, OR = 0.99, $p = 0.488$), as was the effect of L2 cumulative language exposure ($b = 0.10$, $SE = 0.05$, OR = 1.10, $p = 0.709$). The lowest AIC value (indicative of the most parsimonious model) was obtained for Model 3, with sociodemographic variables, non-verbal IQ and L1 and L2 word comprehension as predictors (see Table S8). The second best model, Model 1, showed a difference in AIC above 4 and below 7 (ΔAIC = 5.98), which yields considerably less support for the possibility that this model could be the best approximating model in the candidate set. Other models yielded a difference in AIC above 7, providing little support for them being the best approximating models (see Burnham & Anderson, 2004 for rules-of-thumb for ΔAIC).

Second-order false-beliefs index

The base model, which included the sociodemographic variables and non-verbal IQ as predictors, was run for the second-order false-beliefs index, but it revealed no significant effects. Model 2 showed significant effects of age ($b = 0.00$, $SE = 0.01$, OR = 1.00, $p = 0.047$), of SES ($b = 0.03$, $SE = 0.13$, OR = 1.32, $p = 0.027$), and of the L1 vocabulary comprehension index ($b = 0.05$, $SE = 0.02$, OR = 1.05, $p = 0.022$). Model 3 revealed the same pattern as Model 2 regarding age, SES and L1 word comprehension index, but the effect of L2 word comprehension was non-significant ($b = 0.01$, $SE = 0.02$, OR = 1.01, $p = 0.475$). Model 4 revealed a significant effect of length of L2 exposure, $b = 0.04$, $SE = 0.02$, OR = 1.04, $p = 0.044$. Model 5 revealed no significant effect of cumulative L1 language exposure, ($b = -0.01$, $SE = 0.01$, OR = 0.99, $p = 0.095$). In Model 6, neither the effects of L1 ($b = -0.01$, $SE = 0.01$, OR = 0.99, $p = 0.135$) nor L2 cumulative language exposure ($b = 0.00$, $SE = 0.01$, OR = 1.00, $p = 0.599$) were significant. The lowest AIC value (indicative of the most parsimonious model) was obtained for Model 2, in which sociodemographic variables (age and SES) and L1 word comprehension index were significant predictors. The second best model, Model 3, showed a small difference in AIC relative to the best model (ΔAIC = 1.47), giving substantial evidence that this model could be alternatively the best approximating model. Model 4 and 5 also showed small differences in AIC relative to the best model (Model 4: ΔAIC = 2.5, Model 5: ΔAIC = 3.77), and the remaining models (Model 1 and Model 6), with ΔAIC between 4 and 7, provided considerably less support for them being the best approximating models. Details of the full models are provided in Table S8.

Discussion

The goal of the current research was to explore the potential differences in ToM between bilinguals and monolinguals aged 4–6. We contrasted a group of Polish–English sequential bilinguals (Polish migrants to the UK) with a group of monolingual peers living in Poland. Importantly, we made all efforts to carefully match the compared groups on several factors that have been previously established as predictors of ToM in monolinguals: age, gender, SES, IQ and L1 word comprehension. Still, perfect matching of the two samples on L1 skills turned out to be impossible, so we used individual children's scores in L1 word comprehension as a covariate in our analyses. The results reveal a new and intricate picture of the role that language proficiency plays in both L1 and L2 in ToM in bilinguals.

Results summary

For monolinguals (tested here as a reference group), we replicated the results of the previous studies: age and language proficiency matter for ToM, and these two variables override the effects of SES on ToM. When we compared bilinguals' and monolinguals' ToM abilities using standard frequentist analysis and Bayesian inference, we found no differences in three of the four indices of ToM: the overall accuracy index and the first- and second-order false-beliefs indices. In other words, we revealed no bilingual advantage for the standard measures of ToM. As such, our results are in line with those of Han and Lee (2013), Kyuchukov and de Villiers (2009), Pearson (2013, study 4), Gordon (2016), and Dahlgren et al. (2017), all of whom found no differences between monolinguals and bilinguals in various

ToM tasks. Nevertheless, a complex and informative pattern of interactions was observed in our more nuanced follow-up analyses.

The frequentist analysis showed a significant group difference (medium effect size) for the overall justification index in TRT, which taps into more demanding ToM ability. As a reminder, the "why" question was asked after the child answered the standard test question. The 'why' question is considered as more demanding as it requires reasoning about the previous answer and verbalizing the reasoning process. Therefore, based on the frequentist analysis, bilinguals presented justification for their ToM reasoning with more ease than monolinguals; however, the Bayes Factor did not provide enough evidence to claim this hypothesis to be true. Importantly, L1 proficiency turned out to be a significant covariate in the regression analysis and its effect size was large. Thus, we conclude that the bilingual advantage in ToM reasoning is related to the language abilities of sequential bilinguals in their native language, which was also the language of testing.

Third, we found that the overall accuracy of ToM ability in bilinguals was best predicted by the model in which only age and word comprehension in L1 were significant predictors. This model explained over 34% of variance in the ToM accuracy score (SES, gender and non-verbal IQ were non-significant). Thus, it is clear that in bilinguals (as in monolinguals) age and L1 proficiency are important for ToM performance (accuracy) in standard tasks (see Astington & Baird, 2005; Milligan et al., 2007). For monolinguals, not only auditory word comprehension but also IQ was a significant predictor. This indicates that ToM development is associated not only with language proficiency but also with fluid intelligence.

Fourth, and most interestingly, in bilinguals the overall ToM justification score was largely (45% of variance) explained by the base model, i.e., the model with sociodemographic and cognitive variables (age, SES, gender, and non-verbal IQ) and L1 word comprehension, where only age, IQ and L1 word comprehension were significant predictors. However, when L2 proficiency was added to the model, 54% of variance in ToM reasoning was explained, and only three predictors remained significant: age, L1 word comprehension and L2 word comprehension. In other words, for questions that were more cognitively and linguistically demanding ("why") than the standard ToM question ("where"), proficiency in both L1 and L2 were significant, despite the fact that only L1 was overtly used as the language of testing.

Finally, as for the other investigated factors related to language experience, neither the length of L2 exposure nor cumulative language exposure to L1 and L2 provided any additional predictive value for the outcomes of ToM tasks. This might be a consequence of the fact that both these measures rely on parental reports, which might lack sufficient sensitivity and validity (see Hansen et al., 2019).

Getting back to our main hypotheses, based on the current results we do not have sufficient grounds to claim that Polish–English bilinguals aged 4–6 have an overall advantage over their monolingual peers in basic ToM abilities. We also did not observe any clear benefits of greater input in L1 and L2 for the ToM abilities. However, our results indicate that in bilinguals, proficiency in both L1 and L2 (as assessed by vocabulary tests) relates to advanced ToM abilities, i.e., ability to verbally express reasoning behind ToM judgments. Such a relation was observed even though the ToM task did not require the L2 use.

Theoretical and methodological implications

Our study is one of the first to directly investigate the impact of language abilities and language input on ToM in bilinguals. Although we did not find support for the idea that L2 exposure plays a role in ToM development in bilinguals, our results paint a more nuanced picture of the interaction between ToM abilities and specific language factors than previously reported. These findings are in line with the conclusions formulated by Gordon (2016) as they highlight that ToM in bilinguals benefits from high proficiency across two languages. However, Gordon observed this relation for standard ToM tasks, whereas our results point to a similar relation in more complex and also second-order belief tasks (see also Buac & Kaushanskaya, 2020 for a similar attempt with older children).

We also provide the first evidence that bilingualism could be related to the enhanced ability to reflect on the mental states of others. This was demonstrated by the bilingual advantage in reasoning about assumed thoughts of characters in various stories. Based on our findings, it appears that in four- to six-year-old children, differences between bilinguals and monolinguals may ONLY be apparent when a challenging ToM task is used. In bilinguals, answers to such challenging questions seem to be dependent not only on children's proficiency in the language of testing but also on their proficiency in the other (nontarget) language. As such, our findings indicate that for some aspects of ToM to be enhanced it may not be enough to have only limited L2 experience (as suggested by Fan et al., 2015). At least for more cognitively demanding ToM abilities, the achieved proficiency in both languages may matter substantially. Why does proficiency in both languages of a bilingual impact ToM reasoning, regardless of the language of testing?

It is still unclear what the exact mechanism that drives the observed effect is. It could be that knowledge of more than one language supports or scaffolds the ToM abilities involved in reasoning about mental states (required when answering difficult "why" questions). It is also possible that the benefit is linked to bilinguals' training in more linguistically demanding situations – for example, switching between their languages when talking with different people (e.g., parents vs teachers or peers in daycare). This experience of adjusting the language to the interlocutor may enhance children's socio-linguistic abilities, but it could also lead to the training of executive functions, which then reciprocally feeds into the ToM advantage. Additionally, being immersed in a second-language environment may stimulate bilinguals to reflect on language as a tool that people use to communicate. Finally, it could be that language proficiency is just a proxy of the intensity and length of L2 learning and immersion. In fact, the impact of language proficiency could also be mediated by social-pragmatic skills or meta awareness, both of which go hand in hand with increasing proficiency. Future research should attempt to tackle the issue of the underlying mechanisms (see Yu et al., 2021 for suggestions of some promising research avenues).

We believe that the finding that L1 and L2 skills relate to advanced ToM in bilinguals opens a new window to investigate the emergence vs. expression hypothesis (Moses, 2001). Future research, testing both simultaneous and sequential bilinguals (with different age of L2 acquisition), could address the critical question beyond the scope of the hypothesis – namely, whether language skill (or input) is FUNDAMENTAL for ToM development (as in emergence hypothesis) or is NEEDED ONLY FOR ITS EXPRESSION (as in the expression hypothesis). In general the complexity of the bilingualism phenomenon seems to be a promising window to address this difficult and broad research question. Moreover, combining precise measurements of L2 exposure (accounting for both its quantity and quality) with the longitudinal design could help discover which aspects of language experience are crucial for the enhanced development of the advanced ToM reasoning observed in bilinguals.

We believe that the current results contribute not only to research on bilingualism but also to broader theorizing about the development of Theory of Mind across the lifespan (Warnell & Redcay, 2019). Apperly (2011) proposed a dual-system theory of mindreading abilities, according to which children might be undergoing an important developmental change after the age of 6 years: they gradually begin using high-level mindreading, which is reflected in earlier performance in standard ToM tasks (in addition to the ceiling effects in accuracy). We extend this proposal by suggesting that bilingualism is an example of an experience that supports the transition from low-level (automatic and efficient) mindreading to high-level mindreading (effortful and flexible), as described by |Apperly (2011, 2021). The use of more than one language on a regular basis is typically grounded in complex social interactions that require effortful monitoring of the language being used by a given interlocutor and selection of the right language in response. As such, bilingualism may constitute a natural context for both training the high-level cognitive abilities that underlie mindreading and for learning to make inferences about other people's minds. Our current findings suggest that bilingual children may indeed manifest these high-level abilities earlier in their development. Importantly, to detect these skills we need to use age-sensitive (more challenging) tasks in which children are asked not only to provide the right solution but also to justify this solution; this requires a child to reflect on their own thinking processes and demonstrate their reasoning about a given social situation. If combined with further qualitative and linguistic analyses of their responses (e.g., the presence of mental terms used by parents – Tompkins et al., 2019; the ability to produce structures containing a complement – Hollebrandse et al., 2014), we could gain a better understanding of the possible mechanisms underlying the development of advanced ToM and extend our understanding of the very nature of ToM. As indicated by Apperly (2011), remarkably little attention has been devoted to children's ability to reflect on the causes, consequences and justifications of people's beliefs – in other words, the study of children's 'folk epistemology'. We would like to encourage researchers to include justification questions when studying ToM, especially in older children.

Limitations and future studies

It should be noted that our study focused solely on migrant bilinguals. Moreover, our bilingual sample was relatively homogenous not only in terms of the type of bilingualism (sequential) but also in terms of L1 dominance. All bilingual children in our sample performed ToM tasks in their dominant language (L1), which assured that the potential effect of poor language skills in the language of the testing of ToM performance was minimized. However, this sample homogeneity may also mean less generalizability to other types of bilinguals. Our findings support the idea that specific types of bilingual experience likely play a crucial role in the formation of ToM abilities in bilingual children. Given the great heterogeneity of the bilingualism phenomenon, it is critical to investigate ToM across different bilingual communities and populations whose language experiences are varied.

Importantly, our results imply that bilingualism not only compensates for weaker skills in the language of testing that are important for ToM, but also enhances advanced ToM development more than language abilities per se. While selecting the control group of monolingual children, we deliberately selected children who had relatively weak(er) skills in L1 in order to make the group more comparable to bilinguals. It cannot be ruled out, however, that if we allowed monolingual children with strong L1, their performance on ToM reasoning abilities would be better than that of bilinguals.

Another unique aspect of our study is that it focused on bilingual children who were children of migrant families and were tested in L1 but not in L2 environment. The monolingual group was tested in their home country – Poland. Notably, out of the 13 studies presented in our review, only one (Goetz, 2003) compared groups which were settled in different environments. Although we ensured that the two groups did not differ in SES, it is currently unknown how the difference in the testing environment (L1 vs. L2) may have impacted the pattern of results.

Finally, it should be noted that although we made an attempt to account for individual differences in non-linguistic abilities (by including in the models participants' scores in fluid intelligence), our analyses did not include additional predictors related to executive control or working memory. As indicated in the Introduction, some accounts of ToM development suggest a crucial, possibly mediating role of EF on ToM development, especially in bilinguals. Therefore, future research should definitely employ not only language-related predictors of ToM, but also EF and working memory.

Conclusions

Our study highlights the role of language skills in ToM development. Although bilinguals did not differ from monolinguals in response accuracy in ToM tasks, they demonstrated better reasoning abilities when providing justification for their ToM responses. Moreover, while the ToM accuracy scores were best predicted by L1 proficiency, the justification scores were best predicted by both L1 and L2 proficiency, even though only L1 was needed to perform the task. Overall, the results paint a more nuanced picture of the impact of bilingualism on ToM development. Learning two languages, even sequentially, likely provides fertile ground for the development of more advanced ToM in children aged 4–6 and making inferences about the mental states of others.

Supplementary Material. The supplementary material for this article can be found at http://doi.org/10.1017/S1366728923000652.

Data Availability Statement. The data that support the findings of this study are openly available in the Open Science Framework repository (OSF) at https://osf.io/pzcxt/

Acknowledgements. We express our gratitude to all the children and parents who participated in the study, as well as to all the Bi-SLIPL team members who contributed to collecting and coding the data. Special thanks to Alba Casado for her work on the first draft of Table 1, and to Agnieszka Dynak, Joanna Kolak-Rodis, Magdalena Łuniewska-Etenkowska and Jakub Szewczyk for creating a set of indices out of the parental questionnaire (Kwestionariusz Rozwoju Językowego). We also thank Mike Timberlake for proofreading the text.

Funding. The research presented in this paper was conducted within the Bi-SLI-Poland project "Cognitive and language development of Polish bilingual children at school entrance age – risks and opportunities". The project was carried out at the Faculty of Psychology, University of Warsaw, Poland in collaboration with the Institute of Psychology, Jagiellonian University, Poland. The project was supported by the Polish Ministry of Science and Higher Education / National Science Centre (Decision 809/N-COST/2010/0). Data collection, data coding and maintenance were also partly supported by a Foundation for Polish Science subsidy to Zofia Wodniecka. The project is linked to the European COST Action IS0804.

Competing interest. The authors declare none.

Notes

[1] Note that the effects of executive demands on verbal ToM tasks may be difficult to separate from the effects of language demands of ToM task (for example, shortening the stories in ToM tasks decreases both the linguistic and executive demands of the tasks, meaning lower demands on working memory and language comprehension). Moreover, especially in the case of EF, different components of EF (e.g., working memory, inhibition, flexibility, planning) may impact different ToM tasks

[2] We are aware of differences between the Polish and English formal educational systems: in England, four and five-year-olds attend the reception and the compulsory classes in schools; however, in Poland, six-year-olds attend introductory classes in preschool and start formal education at schools after their seventh birthday. However, technically, before the age of 7, both Polish and English children are rarely fluent in writing and reading in their native language, thus their metalinguistic skills are not mature.

[3] Compared to Schroeder's (2018) and Yu et al.'s (2021) meta-analysis, we took into account only studies that 1) were published; 2) studied children not older than 6 years, i.e., before they start to use language for learning; 3) directly compared bilinguals' and monolinguals' results in ToM tasks.

[4] Ninety-one is the maximum score for a child's exposure in a given language. A score of 91 for Polish presupposes that when living in Poland the children had the maximum exposure to Polish (i.e., 91) and none to English. This might be an oversimplified view, as some children could have had some (possibly irregular) exposure to English while still living in Poland.

[5] (time spent in Poland) * 0 denotes no (zero) exposure to English while living in Poland. This might be an oversimplified view, as some children could have had some (possibly irregular) exposure to English while still living in Poland.

[6] The Bayes factor is the ratio of the likelihood of one particular hypothesis (e.g., the alternative, BF_{10}) to the likelihood of another hypothesis (e.g., null hypothesis, BF_{01}) given the observed data. Thus, the Bayes Factor can quantify the support for one model (BF_{10} or BF_{01}) over another, thus amending the flaw of frequentist inference. A Bayes Factor of around 1 means there is no evidence for one hypothesis over another; a BF of 1–3 means very weak or anecdotal evidence; a BF of 3–10 means moderate evidence; a BF of 10–30 means strong evidence; and a BF of 30–100 and above means very strong/extreme evidence for one hypothesis over another (see Schmalz et al., 2021).

References

Apperly, I. (2011). *Mindreading. The cognitive basis of "theory of mind".* Psychology Press. https://doi.org/10.4324/9780203833926

Apperly, I. (2021). Cognitive basis of mindreading in middle childhood and adolescence. in R. T. Devine & S. Lecce (eds), *Theory of Mind in Middle Childhood and Adolescence: Integrating Multiple Perspectives.* 1st ed, Routledge, pp. 37–54. https://doi.org/10.4324/9780429326899-4

Astington, J. W., & Baird, J. A. (2005). *Why Language Matters for Theory of Mind.* Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195159912.001.0001

Astington, J. W., & Hughes, C. (2013). Theory of mind: Self-reflection and social understanding. In P. D. Zelazo (Ed.) *The Oxford Handbook of Developmental Psychology.* Vol 2 (Self and Other). New York: Oxford University Press, pp. 398–424.

Białecka-Pikul, M., Szpak, M., Haman, E., & Mieszkowska, K. (2018). Teoria umysłu i jej pomiar u dzieci w wieku 4-6 lat: Test Refleksji nad Myśleniem [Theory of Mind and its Measurement in Children from 4 to 6 Years of Age: Reflection on Thinking Test]. *Psychologia Rozwojowa [Developmental Psychology]*, *23*(1), 41–68.8.

Bialystok, E., & Craik, F. I. (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., & Senman, L. (2004). Executive processes in appearance–reality tasks: The role of inhibition of attention and symbolic representation. *Child Development*, *75*(2), 562–579. https://doi.org/10.1111/j.1467-8624.2004.00693.x

Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary differences in monolingual and bilingual children. *Bilingualism* (*Cambridge, England*), *13*(4), 525–531. https://doi.org/10.1017/S1366728909990423

Bonifacci, P., Barbieri, M., Tomassini, M., & Roch, M. (2017). In few words: Linguistic gap but adequate narrative structure in preschool bilingual children. *Journal of Child Language*, 1–28. https://doi.org/10.1017/S0305000917000149

Bronwell, R. (2000). *Expressive One-Word Picture Vocabulary Test* (EOWPVT).

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

Burnham, K. P., & Anderson, D. R. (2004). Multimodel Inference: Understanding AIC and BIC in Model Selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Call, J., & Tomasello, M. (1999). A nonverbal false belief task: The performance of children and great apes. *Child Development*, *70*(2), 381–395. https://doi.org/10.1111/1467-8624.00028

Carpendale, J. I., & Chandler, M. J. (1996). On the Distinction between False Belief Understanding and Subscribing to an Interpretive Theory of Mind. *Child Development*, *67*(4), 1686–1706. https://doi.org/10.1111/j.1467-8624.1996.tb01821.x

Chandler, M., & Lalonde, C. (1996). Shifting to an interpretive theory of mind: 5- to 7-year-olds' changing conceptions of mental life. In: A. J. Sameroff & M. M. Haith (Eds.) *The Five to Seven Year Shift: The Age of Reason and Responsibility*. Chicago, IL US: University of Chicago Press, pp. 11–139.

Dahlgren, S., Almén, H., & Dahlgren Sandberg, A. (2017). Theory of mind and executive functions in young bilingual children. *Journal of Genetic Psychology*, *178*, 303–307. https://doi.org/10.1080/00221325.2017.1361376

de Villiers, J. G. (1991). Why Questions? *University of Massachusetts Occasional Papers in Linguistics*, *17*(1). https://scholarworks.umass.edu/umop/vol17/iss1/8

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

de Villiers, J. G., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development*, *17*(1), 1037–1060. https://doi.org/10.1016/S0885-2014(02)00073-4

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

Devine, R. T., & Hughes, C. (2018). Family correlates of false belief understanding in early childhood: A meta-analysis. *Child Development*, *89*, 971–987. https://doi.org/10.1111/cdev.12682

Diaz, V., & Farrar, M. J. (2018a). 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

Diaz, V., & Farrar, M. J. (2018b). 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

Doherty, M. J. (2000). Children's understanding of homonymy: Metalinguistic awareness and false belief. *Journal of Child Language*, *27*, 367–392. https://doi.org/10.1017/S0305000900004153

Dunn, L. M., & Dunn, L. M. (1981). *Peabody Picture Vocabulary Test – Revised Form*. Toronto, Canada: Psycan.

Dunn, L. M., & Dunn, L. M. (1997). *Peabody Picture Vocabulary Test* (3rd ed.). Circle Pines, MN: American Guidance Service

Dunn, L. M., & Dunn, L. M. (2007). *Peabody Picture Vocabulary Test* (4th ed.). Circle Pines, MN: American Guidance Service.

Dunn, L. M., Dunn, J., & Styles, B. (2009). *British Picture Vocabulary Scale* (BPVS3). GL/Assessment.

Fan, S. P., Liberman, Z., Keysar, B., & Kinzler, K. D. (2015). The exposure advantage: early exposure to a multilingual environment promotes effective communication. *Psychological Science*, *26*, 1090–1097. https://doi.org/10.1177/0956797615574699

Farhadian, M., Abdullah, R., Mansor, M., Redzuan, M., Gazanizadand, N., & Kumar, V. (2010). Theory of mind in bilingual and monolingual preschool children. *Journal of Psychology*, *1*, 39–46. https://doi.org/10.1080/09764224.2010.11885444

Gardner, M. (1985). ROWPVT: *Receptive One-word Picture Vocabulary Test*. Novato, CA: Academic Therapy Publications.

Gardner, M. (2000). *Receptive One-Word Picture Vocabulary Test*. Novato, CA: Academic Therapy Publications

Genesee, F., Boivin, I., & Nicoladis, E. (1996). Talking with strangers: A study of bilingual children's communicative competence. *Applied Psycholinguistics*, *17*, 427–442. https://doi.org/10.1017/S0142716400008183

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

Gordon, K. R. (2016). High proficiency across two languages is related to better mental state reasoning for bilingual children. *Journal of Child Language*, *43*, 407–424. https://doi.org/10.1017/S0305000915000276

Hadwin, J., & Perner, J. (1991). Pleased and surprised: Children's cognitive theory of emotion. *British Journal of Developmental Psychology*, *9*(2), 215–234. https://doi.org/10.1111/j.2044-835X.1991.tb00872.x

Haman, E., & Fronczyk, K. (2012). *Obrazkowy Test Słownikowy – Rozumienie (OTSR)*. Gdańsk: Pracownia Testów Psychologicznych i Pedagogicznych.

Haman, E., Wodniecka, Z., Marecka, M., Szewczyk, J., Białecka-Pikul, M., Otwinowska, A., Mieszkowska, K., Łuniewska, M., Kołak, J., Miękisz, A., Kacprzak, A., Banasik, N., & Foryś-Nogala, M. (2017). How Does L1 and L2 Exposure Impact L1 Performance in Bilingual Children? Evidence from Polish–English Migrants to the United Kingdom. *Frontiers in Psychology*, *8*, 1444. https://doi.org/10.3389/fpsyg.2017.01444

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

Hansen, P., Łuniewska, M., Simonsen, H. G., Haman, E., Mieszkowska, K., Kołak, J., & Wodniecka, Z. (2019). Picture-based vocabulary assessment versus parental questionnaires: A cross-linguistic study of bilingual assessment methods. *International Journal of Bilingualism*, *23*(2), 437–456. https://doi.org/10.1177/1367006917733067

Hoff, E., Rumiche, R., Burridge, A., Ribot, K. M., & Welsh, S. N. (2014). Expressive vocabulary development in children from bilingual and monolingual homes: A longitudinal study from two to four years. *Early Childhood Research Quarterly*, *29*(4), 433–444. https://doi.org/10.1016/j.ecresq.2014.04.012

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods*, *24*(5), 539–556. https://doi.org/10.1037/met0000201

Hollebrandse, B., van Hout, A., & Hendriks, P. (2014). Children's first and second-order false-belief reasoning in a verbal and a low-verbal task. *Synthese*, *191*(3), 321–333. https://doi.org/10.1007/s11229-012-0169-9

Hughes, C., & Devine, R. (2019). Learning to read minds. A synthesis of social and cognitive perspective. In D. Whitebread, V. Grau, K. Kumpuleinen, M.M. McClelland, N. E. Perry & D. Pino-Paternak, (Eds.) *The SAGE Handbook of Developmental Psychology and Early Childhood Education*, Sage, pp. 169–184. http://dx.doi.org/10.4135/9781526470393

Hughes, C., Adlam, A., Happé, F., Jackson, J., Taylor, A., & Caspi, A. (2000). Good test–retest reliability for standard and advanced false-belief tasks across a wide range of abilities. *Journal of Child Psychology and Psychiatry*, *41*, 483–490. https://doi.org/10.1111/1469-7610.00633

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

Jaworowska, A., & Szustrowa, T. (2003). *Test Matryc Ravena – Wersja Kolorowa [Rave's Matrix Test- coloured version]*. Warsaw: Pracownia Testów Psychologicznych PTP.

Kim, Y. T., Chang, H. S., Lim, S. S., & Baek, H. J. (1995). *Korean Picture Vocabulary Test*. Seoul: Seoul Community Rehabilitation Center

Kovács, Á. 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

Kurdek, L. A., & Rodgon, M. M. (1975). Perceptual, cognitive and affective perspective-taking in kindergarten through sixth-grade children. *Developmental Psychology*, *11*, 643–650. https://doi.org/10.1037/0012-1649.11.5.643

Kuś, K., Otwinowska, A., Banasik, N., & Kiebzak-Mandera, D. (2012). Kwestionariusz Rozwoju Językowego (Language Development Questionnaire). Polish adaptation of the 1st version of PaBiQ, developed within COST Action IS0804. Unpublished material. Faculty of Psychology, University of Warsaw.

Kyuchukov, H., & de Villiers, J. (2009). Theory of mind and evidentiality in romani-bulgarian bilingual children. *Psychology of Language and Communication*, *13*, 21–34. https://doi.org/10.2478/v10057-009-0007-4

Lalonde, C. E., & Chandler, M. J. (2002). Children's understanding of interpretation. *New Ideas in Psychology*, *20*(2), 163–198. https://doi.org/10.1016/S0732-118X(02)00007-7

Lewis, C. (1994). Episodes, events, and narratives in the child's understanding of mind. In: C. Lewis & P. Mitchell (Eds.), *Children's Early Understanding of Mind: Origins and Development*. Hove, UK: Erlbaum, pp. 457–480.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false-belief understanding: A training study. *Child Development*, *74*(4), 1130– 1144. https://doi.org/10.1111/1467-8624.00597

Lombrozo, T. (2006). The structure and function of explanations. *Trends in Cognitive Sciences*, *10*(10), 464–470. https://doi.org/10.1016/j.tics.2006.08.004

Łuniewska, M., Wójcik, M., Kołak, J., Mieszkowska, K., Wodniecka, Z., & Haman, E. (2022). Word knowledge and lexical access in monolingual and bilingual migrant children: Impact of word properties. *Language Acquisition*, *29*(2), 135-164. https://doi.org/10.1080/10489223.2021.1973475

Mazza, M., Mariano, M., Peretti, S., Masedu, F., Pino, M. C., & Valenti, M. (2017). The role of theory of mind on social information processing in children with autism spectrum disorders: A mediation analysis. *Journal of autism and developmental disorders*, *47*(5), 1369–1379. https://doi.org/10.1007/s10803-017-3069-5

McCarthy, D. (1972). *McCarthy Scales of Children's Abilities*. New York: The Psychological Corporation.

Miller, S. A. (2012). *Theory of mind: Beyond the preschool years*. Psychology Press.

Milligan, K., Astington, J. W., & Dack, L. A. (2007). Language and theory of mind: Meta-analysis of the relation between language ability and false-belief understanding. *Child Development*, *78*(2), 622–646. https://doi.org/10.1111/j.1467-8624.2007.01018.x

Moses, L. J. (2001). Executive accounts of theory-of-mind development. *Child Development*, *72*(3), 688-690. https://doi.org/10.1111/1467-8624.00306

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

Pearson, D. K. (2013). *Effect of Language Background on Metalinguistic Awareness and Theory of Mind*. Unpublished doctoral dissertation. University of Stirling

Perner, J., & Wimmer, H. (1985). "John thinks that Mary thinks that…" attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, *39*(3), 437–471. https://doi.org/10.1016/0022-0965(85)90051-7

Perner, J., Leekam, S. R., & Wimmer, H. (1987). Three-year-olds' difficulty with false belief: The case for a conceptual deficit. *British Journal of Developmental Psychology*, *5*(2), 125–137. https://doi.org/10.1111/j.2044-835X.1987.tb01048.x

https://doi.org/10.1017/S1366728923000652 Published online by Cambridge University Press

Rubio-Fernández, P. (2017). Why are bilinguals better than monolinguals at false-belief tasks?. *Psychonomic bulletin & review*, *24*(3), 987–998. https://doi.org/10.3758/s13423-016-1143-1

Sabbagh, M. A., & Bowman, L. C. (2018). Theory of Mind. In J. H. Wixed (*Editor-in-Chief*) & S. Ghetti (Ed.) *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience, (Vol. 4, 4th ed.) Developmental & Social Psychology*. New York: Wiley, pp. 249–289.

Schmalz, X., Biurrun Manresa, J., & Zhang, L. (2021). What is a Bayes factor? *Psychological Methods*. https://doi.org/10.1037/met0000421

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

Tager-Flusberg, H., & Sullivan, K. (1994). A second look at second-order belief attribution in autism. *Journal of Autism and Developmental Disorders,* 24, *5*, 577–586. https://doi.org/10.1007/BF02172139

Talwar, V., Gordon, H. M., & Lee, K. (2007). Lying in the elementary school years: Verbal deception and its relation to second-order belief understanding. *Developmental Psychology*, *43*(3), 804–810. https://doi.org/10.1037/0012-1649.43.3.804

Tompkins, V., Farrar, M. J., & Montgomery, D. E. (2019). Speaking Your Mind: Language and Narrative in Young Children's Theory of Mind Development. *Advances in Child Development and Behavior*, *56*, 109–140. https://doi.org/10.106/bs.acdb.2018.11.003

Tuller, L. (2015). Clinical Use of Parental Questionnaires in Multilingual Contexts. In S. Armon-Lotem, J. de Jong, & N. Meir (Eds.), *Assessing Multilingual Children: Disentangling Bilingualism from Language Impairment*. Multilingual Matters.

Vinden, P. G. (1999). Children's understanding of mind and emotion: A multi-culture study. *Cognition & Emotion*, *13*(1), 19–48. https://doi.org/10.1080/026999399379357

Walker, S. (2005). Gender differences in the relationship between young children's peer-related social competence and individual differences in theory of mind. *The Journal of Genetic Psychology*, *166*(3), 297–312. https://doi.org/10.3200/GNTP.166.3.297-312

Warnell, K. R., & Redcay, E. (2019). Minimal coherence among varied theory of mind measures in childhood and adulthood. *Cognition*, *191*, 103997. https://doi.org/ 10.1016/j.cognition.2019.06.009

Wellman, H. M. (2014). *Making Minds: How Theory of Mind Develops*. Oxford: Oxford University Press. https://doi.org/10.1080/15248372.2016.1205337

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

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

Wellman, H. M., Fang, F., & Peterson, C. C. (2011). Sequential Progressions in a Theory-of-Mind Scale: Longitudinal Perspectives. *Child Development*, *82* (3), 780–792. https://doi.org/10.1111/j.1467-8624.2011.01583.x

Wiig, E. H., Secord, W. A., & Semel, E. (2004). *Clinical evaluation of language fundamentals* (2nd ed.). San Antonio, TX: Harcourt Assessment.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition*, *13*(1), 103–128. https://doi.org/10.1016/0010-0277(83)90004-5

Woolfe, T., Want, S. C., & Siegal, M. (2002). Signposts to development: Theory of mind in deaf children. *Child Development*, *73*(3), 768–778. https://doi.org/10.1111/1467-8624.00437

Yow, W. Q., & Markman, E. M. (2015). A bilingual advantage in how children integrate multiple cues to understand a speaker's referential intent. *Bilingualism Language and Cognition*, *18*, 391–399. https://doi.org/10.1017/S1366728914000133

Yu, C. L., Kovelman, I., & Wellman, H. M. (2021). How Bilingualism Informs Theory of Mind Development. *Child Development Perspectives*, *15*(3), 154–159. https://doi.org/10.1111/cdep.12412

https://doi.org/10.1017/S1366728923000652 Published online by Cambridge University Press