Cognitive Development 63 (2022) 101216

Contents lists available at ScienceDirect

Cognitive Development

journal homepage: www.elsevier.com/locate/cogdev

Is cultural variation the norm? A closer look at sequencing of the theory of mind scale

Hande Ilgaz[a,*], Jedediah Wilfred Papas Allen[a], Feride Nur Haskaraca[b]

[a] Psychology Department, Bilkent University, Turkey
[b] Department of Developmental Psychology, University of Göttingen, Germany

ARTICLE INFO

Keywords:
Theory of mind scale
Theory of mind sequence
Intra-cultural differences
Executive functioning
Preschool cognitive development

ABSTRACT

Wellman and Liu's (2004) ToM scale canonized efforts to generate a developmentally nuanced understanding of ToM. Further elaboration has come from studies showing some variability in task sequencing across two broad categories of culture (i.e., 'Collectivist', 'Individualist'). The current study contributes to our understanding of ToM by exploring intra-cultural variation in task sequencing for a Turkish sample. The ToM scale, language, and EF tasks were administered to 366 preschoolers. When analyzed as a single group, preschoolers showed a sequence most consistent with Chinese/Iranian samples. However, when children were grouped according to age, 3-year-olds were most similar to the US/Australian samples, 4-year-olds were most similar to Chinese/Iranian samples, and 5-year-olds showed a new sequence where *knowledge access* was the easiest. The analyzes suggest that EF alone was related to the differences in sequencing. Current findings imply that explaining sequence differences may require considering the interactive effects of culture and cognitive abilities.

1. Introduction

Children's developing ability to understand and coordinate the psychological perspectives of self and others (i.e., Theory of Mind) has classically been measured with false belief tasks. Tasks such as 'change of location' or 'unexpected contents' assess young children's ability to understand that agents may have beliefs that are different from their own (i.e., false beliefs) that motivate their behavior (Wimmer & Perner, 1983). Wellman (2014) has argued that belief understanding is one concept among a series of interrelated concepts about the mind (e.g., desires, emotions). For this reason, he posited that overreliance on false belief tasks to assess children's unfolding development in understanding minds may provide a "limited and misleading" picture (Wellman, 2014, p. 24).

In order to develop a comprehensive measure of children's theory of mind abilities, Wellman and Liu (2004) created a scale based on a meta-analysis of the available literature. This analysis indicated that children's appreciation of diverse desires, diverse beliefs, and the relation between perceptual access and knowledge were all likely to contribute to belief formation. Further, these abilities all develop before children's ability to appreciate the effect of false beliefs on behavior. As a result, a five-task ToM scale was created to assess children's understanding of various mental states in different situations. According to Wellman (2014), the significance of this scale extends beyond methodological and assessment considerations. Inherent in the ToM scale are two theoretical points: (a) mental state concepts can be scaled in terms of conceptual complexity, and (b) there is a universal sequence for children's developmental achievements in understanding these mental state concepts.

* Correspondence to: Bilkent University, Psychology Department, 06800 Ankara, Turkey.
E-mail address: hande.ilgaz@bilkent.edu.tr (H. Ilgaz).

https://doi.org/10.1016/j.cogdev.2022.101216
Received 26 March 2021; Received in revised form 29 June 2022; Accepted 1 July 2022
Available online 29 July 2022

In the past decade, Wellman and colleagues have examined whether the degree of conceptual difficulty on the ToM scale tasks varies by culture (e.g., Etel & Yagmurlu, 2015; Selcuk et al., 2018; Shahaeian et al., 2011; Wellman et al., 2006). These studies have documented two sequences that differ from one another in terms of whether the diverse belief task is passed before or after the knowledge access task. The sequence in which diverse beliefs is mastered before knowledge access has been somewhat imprecisely termed the "Western sequence" since it is consistently observed with children from Western cultures such as the US, Germany, and Australia (Kristen et al., 2006; Peterson et al., 2005; Wellman & Liu, 2004). The reverse sequence has been termed the "Eastern sequence" since it has been observed with children from China and Iran (Shahaeian et al., 2011; Wellman et al., 2006).

Two studies with Turkish children have found evidence for the existence of both sequences (i.e., Western and Eastern). Interestingly, the Western pattern was found for children from disadvantaged backgrounds raised in a government institution (Etel & Yagmurlu, 2015). In contrast, the Eastern sequence was shown for children raised in a typical home environment (Selcuk et al., 2018). In light of these studies that show the presence of both types of ToM sequence in Turkish samples, the current study aimed to add to these efforts by investigating (a) task sequencing with a large middle-class Turkish sample, (b) the possible relations between variations in sequence, age, and cognitive abilities that are known to contribute to ToM achievements (i.e., language, executive functions).

1.1. *Universal sequence vs. cultural diversity*

One of the most intriguing aspects of false belief understanding is the robust performance shift that happens in explicit verbal ToM tasks during the preschool years across cultures (Callaghan et al., 2005; Wellman et al., 2001). Wellman and Liu's scale situates this shift within a broader developmental context involving other mental states that are thought to be conceptually related to false belief understanding. In addition, this scale is proposed to provide evidence for a gradual progression of ToM understanding from "simpler" to "more complex" constructs (Gopnik & Wellman, 1992; Wellman, 1990, 2014). It is important to note that as opposed to the explicit verbal ToM tasks, there is a vast albeit dissonant literature on infants', preschoolers', and adults' performance on implicit ToM tasks, which require spontaneous responding (for reviews, see Dörrenberg et al., 2018; Scott & Baillargeon, 2017). The spontaneous responding tasks have exclusively focused on false belief understanding, and the subcomponents of ToM ability, such as appreciating the diversity of beliefs or desires, have not been studied. For this reason, the current paper does not include a discussion of this literature.

Wellman and Liu (2004; Study 2) administered seven tasks to 75 American children ranging in age from 2;11 to 6;6. Based on a scaling analysis, they kept five of the seven tasks for the scale. Of these five tasks, the simplest was the *diverse desires* task (DD) which taps into children's appreciation that others' desires may differ from one's own. The second task was the *diverse beliefs* task (DB), which assesses whether children appreciate that two agents (e.g., self and other) could hold diverse beliefs about the same situation (e.g., the unknown location of a lost cat). The third task was the *knowledge access* task (KA), in which children had to appreciate the link between perception and knowledge (i.e., someone who has not seen inside a box will not know its contents). The fourth task was the *contents false belief* task (CFB), which required children to understand that a familiar box (e.g., a band-aid box) that holds unexpected contents (e.g., a crayon) would lead an unsuspecting other to assume the usual contents (i.e., band-aids). The *Explicit False Belief* (EFB) task, which was fashioned after the classic change of location task, was also used but was not included in the scaling analysis. This was because both EFB and CFB measured the same ability (i.e., false belief) and appeared as the fourth most difficult task in the scale. The most complex task in this scale was the *apparent/real emotions* task (ARE), which asks children to appreciate that agents sometimes display different emotions than what they feel (e.g., appearing happy when actually sad).

Studies with Australian children have shown the same sequence of complexity as the original study with American children (e.g., Peterson et al., 2005), whereas studies with Chinese and Iranian children have shown a slight difference in the pattern of passing the tasks (Shahaeian et al., 2011; Wellman et al., 2006). Specifically, children from China and Iran pass the knowledge access task earlier than the diverse belief task. In their study with Chinese children, Wellman et al. (2006) analyzed whether children showed significantly different performance in pairs of tasks. Their results showed that the diverse belief task was indeed more difficult for children than the knowledge access task.

Another comparison of note is between DD and KA tasks. American and Australian children have generally performed at ceiling in the DD task (i.e., 95 %), while their KA performance has been subject to development (73% in American, 82 % in Australian samples). However, a review of children's performance of DD in other cultures does not indicate a ceiling effect (i.e., 82–91 %). In fact, when DD

Table 1
The percentage of children who passed each ToM task by study.

Tasks	Wellman and Liu (2004)	Peterson et al. (2005)	Wellman et al. (2006)	Shahaeian et al. (2011)	Etel and Yagmurlu (2015)	Selcuk et al. (2018)
	American children (%)	Australian children (%)	Chinese children (%)	Iranian children (%)	Turkish children (%)	
DD	95	95	89	86	91	82
DB	84	85	71[E]	47[E]	71	64[E]
KA	73	82	79[E]	88[E]	44	79[E]
CFB	59	32	54	16	12	24
ARE	32	19	37	17	19	19

Note: DD: Diverse Desire, DB: Diverse Belief, KA: Knowledge Access, CFB: Contents False Belief, ARE: Apparent Real Emotions.
[E] Indicates the Eastern patterns where KA is easier than DB.

and KA performances are compared, they are either very close (i.e., Turkish middle-class 82 % vs. 79 % in Selcuk et al., 2018) or reveal the reverse pattern of performance (86 % vs. 88 % in Shahaeian et al., 2011; 86 % vs. 94 % in Shahaeian et al., 2014). These studies do not include pairwise comparisons to investigate whether performance on these tasks was different. However, these raw results hint at the possibility of other orders. Table 1

The reasons for the documented differences between children from the Iranian and Chinese vs. the Australian and American cultures are not clear. In his review of ToM research, Wellman (2014) highlights commonalities between the two cultures that may similarly shape patterns of ToM performance. Against a backdrop of differences (e.g., Indo-European vs. Sino-Tibetan language families, Muslim vs. Confucian belief systems), the two cultures may share some child-rearing beliefs and practices that emphasize filial respect and conformity. However, whether these suggested commonalities exist and influence children's developing ToM abilities is still an open research question. That is because extant work comparing Eastern cultures with Western cultures has not specifically investigated the effects of the cultural commonalities and differences on children's ToM scale performance (Shahaeian et al., 2011; Shahaeian et al., 2014; Wellman et al., 2006). A recent intra-cultural investigation of Israeli-Arab Muslim mothers and their children provides a relevant starting point. Kabha and Berger (2020) examined the possible relations between variance in parents' individualistic vs. collectivistic values and the sequence of performance on the ToM scale. This study did not find any significant relations between cultural values and ToM performance.

Studies that show sequence differences across cultures (Shahaeian et al., 2011; Shahaeian et al., 2014; Wellman et al., 2006) suggest, in broad strokes, that folk ethnopsychologies of the mind may impact children's socio-cognitive development (Lillard, 1998). Further, the two studies with Turkish samples provide evidence for the possibility that there may be different ethnopsychologies present in the same culture. First, Etel and Yagmurlu (2015) showed that Turkish children displayed the same sequence of ToM performance as observed in the American and Australian samples. However, the sample in that study was atypical in that the children lived under government guardianship either because their biological families could not provide adequate support, were absent from the child's life, or had been suspected of abuse. These children were participating in a recent government initiative in which 10–12 children live with three adult childcare providers in a house. This program aimed to place siblings in the same house to maintain and foster emotional bonds. In short, Etel and Yagmurlu's findings showed that the pattern of sequence characteristic of American and Australian children was also the dominant pattern in Turkish children from extremely disadvantaged backgrounds.

The second study that investigated patterns of ToM performance in Turkish children used a more typical sample (Selcuk et al., 2018). These children came from 5 large cities in Turkey. The cities ranged in population from 1 to 12 million. This study included 260 children and showed that the most prevalent pattern was the Eastern pattern (i.e., Chinese/Iranian, where KA > DB). Importantly, this study also showed that a minority of children ($n = 21$) showed the Western pattern (i.e., American/Australian, where DB > KA). This was the first study to document different patterns of ToM performance within the same culture. Further analysis showed that the number of adults that lived with the children predicted whether they passed the knowledge access task before the diverse belief task or not. That is, those children who lived with more adults were more likely to show the American/Australian pattern.

Taken together, these two studies (Etel & Yagmurlu, 2015; Selcuk et al., 2018) showed that differences in patterns of performance on the ToM scale were observed between subgroups of the same culture (Turkish children from intact family backgrounds vs. those under government guardianship). Additionally, even within the same subgroup, different patterns of performance coexist (Selcuk et al., 2018). It is challenging to make sense of why the number of adults (but not siblings) would affect young children's socio-cognitive understanding. Perhaps in households with more adults, children get more frequent and more extended conversational interactions. However, this remains a research question to be answered.

Another possibility for the differences in patterns of performance on the ToM scale could be child variables, especially cognitive variables that have been consistently reported to be related to ToM abilities. Among these variables, the ones that are consistently correlated with and predictive of ToM ability, cross-sectionally and longitudinally, are language and executive functions (EF) (e.g., Milligan et al., 2007; Moses & Tahiroglu, 2010). The majority of the research that has shown these relations have been carried out with false belief tasks. There is a scarcity of research examining the relations between cognitive variables and children's success on the ToM scale. One of the possible advantages of the scale is its potential to show gradual developments in ToM ability by focusing on its constituent components. More importantly, the investigation of ToM scale sequences in different cultures has the potential to provide evidence for or against arguments about the universality of core folk psychological concepts. It is thus critical to investigate the possible relations between cognitive variables and the sequence of performance as well as the relations between the individual tasks.

Three studies have used the ToM scale and sought to find relations between ToM performance and EF abilities. Henning et al. (2011) investigated the relation between EF abilities (as measured with DCCS) and ToM scale performance for 3-, 4-, 5-, and 6-year-old German children. German children were found to exhibit the Australian/American pattern, and their EF scores (taken as either aggregate or post-switch) were differentially related to the ToM scale tasks. Specifically, the DB, KA, and the false belief tasks (CFB and EFB) were related to children's post-switch EF scores. Duh et al. (2016) investigated the relation between EF and ToM scale performance with a large-scale study that included 997 Chinese children from Chengdu. EF tasks that measured working memory and inhibition predicted children's overall ToM scale performance. A particularly relevant aspect of the results concerned sequence differences. Children in this study showed not only the expected DB and KA reversal but also a reversal of the ARE and EFB tasks. Concerning the difference between their results and those of Wellman et al. (2006), the authors emphasize the diversity of subcultures within China (Beijing vs. Chengdu; rice growers vs. wheat growers; Northern vs. Southern China). Finally, a study with Turkish children has shown that children's KA and FBU (aggregate of CFB and EFB) scores were correlated with EF abilities, as measured by tasks that mainly measure inhibition, both cross-sectionally and with a one-year delay (Doenyas et al., 2018). Furthermore, early EF abilities predicted later ToM total scores even when age, language abilities, and ToM at Time 1 were controlled. Taken together, these studies have shown relations between different aspects of EF (inhibition, rule switching, and working memory) and ToM aggregate scores. However, none of the studies has explored whether EF abilities make a difference in the sequence with which children pass ToM scale tasks.

The relation between language and ToM abilities is extensive and well-established (e.g., Milligan et al., 2007). The only study that reports performance on the individual tasks of the ToM Scale task and language is by Doenyas et al. (2018). Their results showed differential relations between language and scale tasks where only KA, FBU at both testing times, and ARE performance at Time 2 were related to receptive language. The current study aimed to investigate whether cognitive variables that have been documented to be related to ToM performance could also be related to the patterns of performance on the ToM scale.

1.2. *The current study*

The current study investigated the patterns of performance on the ToM scale with a large sample of preschoolers. The data for this study came from six separate studies that all used the ToM scale to measure children's socio-cognitive development. All studies used the same testing protocol and recruited children from schools that serve families with similar socioeconomic backgrounds. The first aim of the study was to see if the reported variability in ToM task sequences could be replicated (Etel & Yagmurlu, 2015; Selcuk et al., 2018).

Based on Selcuk et al. (2018), we expected that the Chinese/Iranian pattern would be the most prevalent. We also expected to see a portion of the sample to show the American/Australian pattern. It is pertinent to note that what we have called 'the American/-Australian pattern' has generally been dubbed the 'Western' or 'Individualistic' pattern in the literature. Similarly, what we have called 'the Chinese/Iranian pattern' has been called the 'Eastern' or 'Collectivistic' pattern. While we appreciate why other researchers have opted for these labels, in our opinion using such terminology substantiates a priori assumptions without direct evidence for their validity. Given that we had no measure that tapped specific societal or cultural attitudes (e.g., collectivism/individualism scale for parents), we opted to refrain from the labels 'Eastern'/'Collectivistic' or 'Western'/'Individualistic'.

Our second aim was to explore whether any of the child variables (i.e., age, EF, language) or environmental variables (i.e., SES, number of siblings) would play a role in determining the pattern of ToM performance. One of the core assumptions regarding the ToM scale is that the tasks are equivalent in extraneous cognitive burden but vary by conceptual complexity. However, it is possible for some tasks in the scale to tax children's cognitive resources more than others. This would mean that early proficiency in a cognitive domain such as executive functions could affect when and which tasks are passed. As a result, variability in cognitive abilities could impact the sequence of performance on the ToM scale. In order to investigate this possibility, we explored whether children's age, EF, and language abilities contributed to the observed patterns of ToM performance.

If we assume that the collectivism vs. individualism dichotomy affects the pattern with which children pass ToM tasks, we may expect SES (an aggregate of parental education and family income) to act as proxy variables in predicting one of the two patterns. While an imperfect indicator, previous work with Turkish samples has shown that mothers with higher education endorse more individualist values, and higher family income is associated with higher individualism in children (Özdikmenli-Demir & Sayıl, 2009). The number of siblings could affect children's ToM development in two ways. One way would be to enhance the richness of children's everyday experiences in conversation, negotiation, and play contexts (Lewis et al., 1996; Perner et al.,1994). However, in the Turkish context, higher number of siblings could also correlate with a more collectivistic parental attitude. Kagitcibasi and Ataca (2005) have shown that, within an urban mother sample, the amount of education differed between mothers from different SES groups and was related to the number of children they had. Urban mothers with fewer years of education were found to have more children as compared to urban mothers with more education. The current study does not have any direct measures that assess parents' cultural attitudes. However, we chose to include SES and the number of siblings due to their suitability as variables that could indicate differences either in children's experiential history or their parents' cultural attitudes. Further, such inclusion would parallel the relevant literature, which has included these variables in work investigating sequence differences with the ToM scale (Henning et al., 2011; Selcuk et al., 2018; Shahaeian, 2015).

2. Method

2.1. *Participants*

Participants were recruited from preschools serving middle-class populations in Ankara (Turkey) between 2013 and 2017. Ankara is the capital of Turkey, with a population of over 5 million. Four hundred eighty children participated in 6 studies that included the ToM scale. Each study was separately reviewed by the University Ethics Board. In addition, appropriate approvals were obtained from the Ministry of Education. Parents signed consent forms for their children's participation. All children verbally assented to the studies. Only participants who completed the scale and were in the desired age range (3- to 5-year-olds) were included in the study. In addition, several children whose age information was not complete were excluded. The total number of eligible participants for the analysis was 366 (51 % girls). The dataset included 132 3-year-olds ($M_{\text{age}} = 41.6$ months, $SD = 2.91$, range = 36–47 months), 119 4-year-olds ($M_{\text{age}} = 53.7$ months, $SD = 3.32$, range = 48–59 months), and 115 5-year-olds ($M_{\text{age}} = 65.2$ months, $SD = 3.3$, range = 60–71 months).

2.2. *Procedure*

The data were collected from six studies that included the Wellman and Liu (2004) ToM scale as the first task in order of administration. Four of these studies (N = 180) collected data on participants' executive functioning abilities through the Dimensional

Change Card Sorting task (DCCS, Zelazo, 2006). In three studies (N = 111), children's receptive and expressive language skills were measured through the Turkish Expressive and Receptive Language Test (TIFALDI, Berument & Güven, 2013). In all of these studies, the tests were administered by undergraduate and graduate students who underwent extensive training. The tests were administered either at preschools (65 %), in the lab (33 %), or at home (2 %).

2.3. Measures

2.3.1. Demographics

Parents reported their family income on a 5-item scale from 1 (less than 1000 TRY) to 5 (more than 7000 TRY). A total of 277 parents had income data available. Overall, participants mainly came from families in which the monthly income was moderate to high. Specifically, 49.5% of these families had an income of 7000 TRY and above, 26.4 % had between 5000 and 7000 TRY, 14.1 % had between 3000 and 5000 TRY, 9.4 % had between 1.000 and 3.000 TRY, and less than 1% had an income less than 1.000 TRY. Ninety percent of the sample reported that they earned more than twice the average minimum wage (i.e., 1335 TRY) in Turkey during these years (TR Ministry of Labor Social Services and Family, 2018).

Mothers' average age was 36.3 years (N = 275, range = 26–54 years) and fathers' average age was 39.6 years (N = 262, range = 26–64 years). Sibling data was available for 288 of the participants (e.g., number of siblings). Half of these participants (50.7 %) were only children, whereas 43.8 % had one sibling and 4.9% had two siblings. Less than 1 % had three siblings. None of the participants had more than three siblings.

Parents reported their highest education level on a 7-item scale from 1 (being illiterate) to 7 (Doctoral degree). Mothers were generally highly educated. Of the 288 mothers who filled out the demographics survey: 85.4 % had university degrees (of these, 19.4 % had master's, and 7.3 % had doctorate degrees), 10.8 % had a high school degree, 3.5 % had a middle school degree or less. Of the fathers, 284 reported information about their education. Fathers were similarly highly educated: 84.6 % of the fathers had university degrees (of these, 18.7 % had master's, and 7.4 % had doctorate degrees), 12.3 % had a high school degree, 3.2 % had a middle school degree or less.

Families' socioeconomic status (SES) was determined by combining monthly income with father and mother education. Since the income and education variables were rated on different scales, participants' scores were converted to z-scores and combined. As a result, 274 families had complete values for the SES variable (M = -0.3, SD = 2.46; range = -9.21 to 4.97).

2.4. Child measures

2.4.1. Theory of Mind

For all six studies, children were tested on the five subtests of the ToM scale (Wellman & Liu, 2004). The subtests were administered in the following order: Diverse Desires (DD), Diverse Beliefs (DB), Knowledge Access (KA), Content False Belief (CFB), and Apparent/Real Emotions (ARE). When necessary, changes were made to some of the objects that appeared in the stories to make them culturally more appropriate (e.g., using a candy box instead of a Band-Aid box). Participants could receive 1 point for each of the five subtests for a total score from 0 to 5.

The DD and DB tasks were very similar in format. In both tasks, participants were presented with a character and two choices. The DD task asked which food would be chosen by a character whose desires (i.e., its likes/dislikes) differed from the participant's, whereas the DB task asked how a character would act given that his/her belief was different from that of the participant. The following two tasks (KA and CFB) were similar in format. In both of these tasks, participants were shown containers with hidden objects inside. The critical difference was that the container was a non-descript box in the KA task with no features that could cue the participants to its contents. This task aimed to measure whether the participant could accurately judge that the character who never saw inside a non-descript box would be ignorant about its contents. In the CFB task, children were shown a familiar container such as a candy box. However, it was revealed that this familiar box had unexpected contents (e.g., a crayon). The CFB task aimed to measure whether the participant could accurately judge the character's false belief. Lastly, the ARE task differed in format from the others since it involved a pre-examination of the participant's knowledge of emotion expressions followed by a short story about a character. This task aimed to measure whether the participant could understand that the character feels one emotion but displays a different one.

2.4.2. Executive Functioning (EF)

In four of the six studies, children's EF abilities were measured through the DCCS task. DCCS was developed by Zelazo (2006) in order to tap into children's inhibition abilities and cognitive flexibility. In the task, children were asked to sort cards, first according to one rule (i.e., color) and then another (i.e., shape). A final phase required children to shift flexibly between the two rules (the border game). Children could get 1 point for each correct sort resulting in a total score of 24.

2.4.3. Turkish Expressive and Receptive Language Test (TIFALDI)

Children's expressive and receptive language skills were measured in three of the six studies using TIFALDI (Berument & Güven, 2013). The task was standardized for Turkish-speaking children between the ages of 2 and 12 and is very similar in its administration to PPVT (Dunn & Dunn, 2007) and EVT (Williams, 2007). In the expressive language subtest, children are shown pictures from various categories (e.g., animals, plants, everyday objects, actions) and are asked to name them. In the receptive language subtest, children are shown four pictures belonging to the same category and are asked to show a particular one. Children's raw scores are calculated based on the number of correct answers they provide, and the standard scores are calculated based on the raw score and the child's age. In this study, we used a composite score for overall language ability, which was formed by combining raw scores in receptive and expressive language subtests. We used raw scores but not standard scores since we already controlled for age in the analyzes when it was relevant.

3. Results

We started our analysis by exploring whether children's ToM performance differed by age and gender. We followed this with other descriptive analyses that focused on the relations between children's performance on the ToM tasks, their scores on tasks that assess related cognitive abilities (i.e., EF, language), and relevant demographic variables (i.e., number of siblings, SES). Following this, we conducted a Guttman analysis to explore whether Turkish children from a middle-SES background showed the patterns of performance previously identified in the literature. Finally, we analyzed the data to see which demographic and cognitive factors contributed to the patterns children displayed.

We first conducted a 3 (age) × 2 (gender) ANOVA to investigate whether children's ToM performance varied by gender or age. We had no specific hypothesis regarding gender. However, we expected that the data would confirm the developmental change in total ToM scores between 3- and 5-year-olds documented in the literature. The analysis confirmed our expectation and revealed a significant main effect of age, F (2, 360) = 36.84, $p < .001$, partial $\eta^2 = .17$. Bonferroni post-hoc tests revealed that 3-year-olds ($M = 2.03$, $SD = 1.00$) differed significantly from both 4- ($M = 2.83$, $SD = 1.16$, $p <.001$) and 5-year-olds ($M = 3.20$, $SD = 1.13$, $p < .001$); and that 4- and 5-year-olds differed from each other, $p < .05$. There was no effect of gender on total ToM score, F (1, 360) = .66, $p = .42$ and there was no interaction of age and gender on ToM scores, F (2, 360) = .08, $p = .93$. Therefore, gender was not included in further analyses.

Next, we explored the relations amongst demographic variables (SES and number of siblings), cognitive variables (EF and language), and ToM performance (see Table 2). In light of the previous literature, we expected moderate to strong correlations between the total ToM score and the cognitive variables when we controlled for age. The relations between SES, number of siblings, and ToM are less clear and less consistent (e.g., Devine & Hughes, 2016; Perner et al., 1994; Peterson & Slaughter, 2003; Yagmurlu et al., 2005). Consequently, we did not have strong predictions regarding the relations between these variables and ToM performance. Consistent with the extant literature (Milligan et al., 2007; Moses & Tahiroglu, 2010), the age-controlled correlation analysis showed that total ToM score was positively correlated with EF (r (177) = .18, p = .02) and language (r (108) = .31, p = .001); however, it was not correlated with SES or number of siblings.

Given the sequence differences for ToM scale performance in the literature, we opted to include the separate ToM tasks in this partial correlation analysis. The results revealed that the two tasks that switch between the "Chinese/Iranian" and "American/Australian" patterns (i.e., KA and DB) were differentially related to the cognitive and demographic variables (see Table 2). Specifically, performance on the diverse belief task was positively related to SES and negatively related to the number of siblings. That is, children from more traditional families with higher numbers of siblings were less likely to pass the diverse belief task, whereas children from more affluent families with more education were more likely to pass the task. In addition, there were positive correlations for performance on the knowledge access task with EF and language. It is important to note that the magnitudes of these correlations were small. Regardless, these analyzes showed that it would be pertinent to include both the cognitive and the demographic variables in further analyzes designed to understand the differences between children who show different ToM patterns of performance.

The proportion of children who passed each ToM task is presented in Table 3. The pattern indicated by the proportions for the total rates of success in each task resembles the pattern found with Chinese, Iranian, and non-disadvantaged Turkish children (Selcuk et al., 2018; Shahaeian et al., 2011; Wellman et al., 2006). To ensure that this sequence indicates progressive difficulty, we ran Guttman scaling analyses on the data (N = 366). If the ToM scale is truly scalable, we would expect children of different age groups to show similar patterns of performance. Accordingly, we first investigated whether there was a prevalent pattern in the overall sample, followed by whether different age groups showed significantly different patterns by administering separate Guttman analyses on subsamples by age.

The logic behind a perfect Guttman Scale is that if an item has a positive score (in this case, the task is passed), all easier items should also have a positive score. In the ideal case, if we know the pattern with which participants pass and fail the tasks and the total

Table 2
Age-controlled correlations between demographic and child variables.

Variables	1	2	3	4	5	6	7	8	9	10
1. DD	–									
2. DB	.23**	–								
3. KA	.00	.01	–							
4. CFB	.05	-.03	.22**	–						
5. ARE	.00	.05	.05	.19**	–					
6. Total ToM	.49**	.55**	.49**	.57**	.46**	–				
7. EF	.12	-.02	.22*	.15*	-.01	.18*	–			
8. Language	-.00	.11	.24*	.17	.24*	.31**	.32**	–		
9. SES	.06	.15*	.05	-.03	-.01	.10	.11	.37**	–	
10. # of Siblings	-.06	-.15*	.03	.10	-.10	-.07	.00	.32**	-.09	–

Note. DD: Diverse Desires, DB: Diverse Beliefs, KA: Knowledge Access, CFB: Content False-Belief, ARE: Apparent/Real Emotion.
*$p < .05$, **$p < .01$ (All significance tests are two-tailed).

Table 3
Turkish children's performance on ToM tasks.

Tasks	3-year-olds	4-year-olds	5-year-olds	Total
Diverse Desires	0.82	0.85	0.77	0.81
Knowledge Access	0.49	0.82	0.89	0.72
Diverse Beliefs	0.52	0.57	0.63	0.57
Content False-Belief	0.17	0.40	0.58	0.37
Apparent/Real Emotion	0.04	0.19	0.33	0.18

score a participant has obtained, we would know which tasks the participant passed or failed. Table 4 shows the Guttman scalogram for the pattern suggested by the number of participants who passed the tasks in the indicated orders.

Next, Green's method was used to calculate the coefficient of reproducibility and index of consistency. The coefficient of reproducibility indicates the confidence with which we can project the items a participant would have passed given their total score, and this number should ideally be equal to or larger than .90. The analysis showed the coefficient of reproducibility to be acceptable at .90. However, we should note that Green's Index of Consistency was .13, which is below the desired level of .50. This index indicates if items are scalable (i.e., can be ordered from easiest to most difficult) (Green, 1956). Another way to check hierarchical complexity is via individual McNemar's χ^2 tests, where each task can only be compared to one other. When we compared each ToM task to its neighbor (in the order indicated in Table 4), we found that all of the tasks were significantly different from their neighbors in the sequence. These analyzes showed that more children passed DD but failed KA as compared to the reverse, McNemar's χ^2 (1) = 8.13, $p < .01$. Similarly, more children passed KA but failed DB as compared to the reverse, McNemar's χ^2 (1) = 17.67, $p < .001$. The analysis also indicated that more children passed DB but failed CFB as compared to the reverse, McNemar's χ^2 (1) = 27.43, $p < .001$. Finally, more children passed CFB but failed ARE as compared to the reverse, McNemar's χ^2 (1) = 41.76, $p < .001$.

Given that previous work with Turkish children has shown evidence for multiple patterns, we decided to specifically investigate if there were children who passed the DB task before the KA task. As expected, of the 100 children who passed only two tasks, 39 passed first DD and next DB, which would be taken as evidence for the 'American/Australian' pattern. As indicated in Table 4, 35 children showed the 'Chinese/Iranian' pattern. There were eight other possible task pairings that the remaining 26 children could show. None of the other patterns approached the rates we observed for the 'Chinese/Iranian' and the 'American/Australian' patterns. Within these 26, the most prevalent pattern was passing the KA and CFB tasks ($n = 10$), followed by passing the DB and KA tasks ($n = 8$). None of the other patterns were shown by four or more children.

As can be viewed in Table 3, children showed different patterns of success on the ToM tasks depending on their age (see Table 3). A Guttman analysis for 3-year-olds ($n = 132$) showed that 64 % fit the pattern in which the DB task was acquired earlier than the KA task (i.e., the 'American/Australian' pattern: DD > DB > KA > CFB > ARE). The coefficient of reproducibility for this analysis was .92, and the index of consistency was .15. In contrast to the 3-year-olds, the pattern in which the KA task is acquired earlier than the DB task (i.e., the 'Chinese/Iranian': DD > KA > DB > CFB > ARE) was the best fit for 4-year-olds (59 %; $n = 119$) with a coefficient of reproducibility of .91 (index of consistency was .13). The most notable finding pertained to the 5-year-olds. More than half of the 5-year-olds (54 %, $n = 115$) fit a new pattern in which the KA task was passed first (New Pattern: KA > DD > DB > CFB > ARE). The coefficient of reproducibility was .90, and the index of consistency was .16.

Next, we conducted a binary logistic regression to investigate whether demographic variables other than age contributed to the type of performance pattern children showed (i.e., KA-before-DB or DB-before-KA). Table 2 showed small correlations between the number of siblings and SES with performance on the DB task. In addition, children's performance on KA was correlated with their language and EF scores. Given that children's performance on these two tasks (i.e., DB, KA) determines which of the two prevalent patterns of performance children show, we decided to include the number of siblings, SES, language, and EF in a logistic regression analysis along with children's age in months (see Table 5).

This logistic regression analysis included only those children who passed *either* the KA or the DB task ($n = 49$). Children who passed or failed *both* tasks were excluded from the analysis. The outcome variable was whether children passed KA before DB (scored as 1) or DB before KA (scored as 2). Multicollinearity among variables was checked according to widely accepted parameters (Bowerman & O'Connell, 1990; Myers, 1990). The VIFs and tolerance statistics for all variables were within the acceptable ranges (i.e., VIF = 1.00–2.30; Tolerance = .43–.97). Model 1 with age as the only predictor variable was significant $\chi^2(1) = 10.88$, $p < .001$. This model

Table 4
Guttman scalogram pattern for a five-item Guttman scale.

Task	Pattern of success (+) and failure (-)					
	1	2	3	4	5	6
Diverse Desires	-	+	+	+	+	+
Knowledge Access	-	-	+	+	+	+
Diverse Beliefs	-	-	-	+	+	+
Contents False-Belief	-	-	-	-	+	+
Apparent/Real Emotion	-	-	-	-	-	+
Number of Children	6	35	35	59	45	21

Table 5
Regression table for predicting models of performance patterns by cognitive and demographic variables.

Variables	B	SE	Exp (B)	Sig.	Nagelkerke R^2	R^2 change
Step 1					.269	+ .269
Age	-.10	.03	.91	.00		
Step 2					.434	.165
Age	-.05	.05	.95	.28		
Number of Siblings	-1.01	.64	.36	.12		
SES	.33	.24	1.39	.18		
Language	-.01	.02	.99	.49		
EF	-.14	.06	.87	.03		

explained 27 % of the variance and correctly classified 71.4 % of the cases. The second model, which included demographic and cognitive variables, in addition to age, was also statistically significant, $\chi^2(5) = 19.01$, $p < .005$. This model explained 43 % of the variance in the pattern and correctly classified 79.6 % of the cases. In this model, the only significant predictor to emerge was EF. Specifically, increases in EF were associated with an increased likelihood of exhibiting the KA-before-DB pattern.

This analysis showed that EF was the most potent concurrent predictor that determined which of the two patterns (i.e., KA-before-DB or DB-before-KA) children would display. In order to ascertain that the other variables' effects were not masked by a correlation between those variables and EF, backwards, forwards, and forced entry methods were also tried. The results were generally consistent in that EF either emerged as the only significant variable (forced entry method) or emerged as one of two significant variables, the other being age (backwards and forwards entry methods). We previously reported Guttman analyses by age group based on the rationale that developmental patterns should remain constant across age. These analyzes had yielded three sequences (i.e., Chinese/Iranian, American/Australian, and a previously unidentified pattern). We did a final follow-up analysis with a series of Guttman analyses based on children's EF performance. Three groups were created according to the percentile to which participants belonged in the distribution of EF scores. Specifically, children whose EF scores fell into the first quartile were categorized as the low EF group, and those in the last quartile were categorized as the high EF group. Children in between constituted the moderate EF group.

These last set of follow-up analyses within EF performance groups mirrored the findings from our analysis within separate age groups. The Guttman analysis on children with lower EF ($n = 50$) showed that 64 % fit the pattern in which the DB task was acquired earlier than the KA task but later than the DD task (i.e., the 'American/Australian' pattern: DD > DB > KA > CFB > ARE). The coefficient of reproducibility for this analysis was.91, and the index of consistency was.15. Children with moderate EF fit the pattern in which the KA task was acquired earlier than the DB task (i.e., the 'Chinese/Iranian' pattern: DD > KA > DB > CFB > ARE). More than half of them (57 %, $n = 89$) followed this pattern with a coefficient of reproducibility of .90 (index of consistency was .13). Children with high EF did not follow either pattern; instead, similar to the pattern previously observed with 5-year-olds, they passed the KA task first before the DD and DB tasks (61 %, $n = 41$: KA > DD > DB > CFB > ARE). The coefficient of reproducibility was.91, and the index of consistency was .16.

4. Discussion

The current study aimed to further explore the reported variability in the literature regarding the sequence in which children pass the ToM scale tasks. Previous research has provided evidence for both inter-cultural (e.g., Iranian vs. Australian, Shahaeian et al., 2011) and intra-cultural variability (e.g., Duh et al., 2016; Selcuk et al., 2018). As stated, the reason for this variability is not clear. Some researchers have speculated that it may be due to broad differences in cultural patterns of relatedness, particularly individualism and collectivism (e.g., Wellman, 2014). Others have emphasized differences in parenting as a function of such cultural attitudes (e.g., Shahaeian et al., 2011; Shahaeian et al., 2014).

Researchers who argue for the effect of individualism vs. collectivism acknowledge that these explanations remain at the level of speculation since they have not measured the parental styles or attitudes of cultural relatedness in relation to children's patterns of success on the ToM scale. One recent exception (i.e., Kabha & Berger, 2020) has failed to show any relation between mother-reported cultural values and Israeli-Arab Muslim children's patterns of ToM scale sequencing. Selcuk et al. (2018) found surprising results with a proxy variable for collectivism vs. individualism (i.e., the number of adults that live in the same house with the child). Their results showed that this variable concurrently predicted the pattern with which children passed the ToM tasks. However, contrary to what would be expected, children who lived with more adults were more likely to show the 'American/Australian' pattern. This finding is difficult to reconcile with a framework that focuses on cultural norms of relatedness. That is, the results seem to show that Turkish children who live in more traditional households (i.e., with extended family) are more likely to show patterns of ToM performance that match children from more individualistic cultures (e.g., Australia or the US).

The current study explored whether cultural variation on the ToM scale was related to cognitive variables that included language and executive functions. The results confirmed a diversity in ToM profiles for a large and relatively homogeneous middle-class sample. Observing this kind of variability in a relatively homogeneous SES sample provides a strong impetus to further investigate the possible effects of cognitive variables on ToM profiles. Perhaps the broadest index of cognitive development is age. In this sense, age can be considered a proxy variable that marks children's developmental levels in major cognitive areas (e.g., language, executive functions) as well as an indication of children's cumulative history of social experience. While age by itself is not a very informative index, it provides us with general developmental trends. A central assumption of the ToM scale is that it measures incremental conceptual complexity in ToM. Hence it is expected that children would show predictable increases in performance by age for each task. From this view, age is not a very interesting variable as the expectation is that older children will pass more levels, but overall, children will display the same general pattern in their success profiles. One of the most intriguing findings of the current study is that 3-, 4-, and 5-year-olds showed different patterns of success on the ToM scale (See Table 3). Most notably, the relative difficulty of the DB and the KA tasks changed in opposite directions with increasing age (i.e., as children got older, KA became easier and DB became harder).

Importantly, these age-related findings are mirrored when the patterns of success on the ToM scale are investigated in relation to children's EF abilities. Specifically, children with lower EF showed the 'American/Australian' pattern, children with moderate EF showed the 'Chinese/Iranian' pattern, and those with more advanced EF showed a previously unidentified pattern where KA was the first task in the sequence. Similarly, logistic regression analysis indicated that EF had a unique effect on the order with which children passed the KA vs. the DB tasks. More importantly, none of the other variables, neither socio-cultural (i.e., number of siblings, SES), cognitive (i.e., language) or general proxy (i.e., age), explained the observed patterns.

4.1. Could confounds explain the results?

Some potential confounds for the sequence differences according to EF could result from a skewed distribution of participants in the different age groups by SES. In other words, we could observe sequence differences if the age groups did not share the same cultural milieu. However, our analysis did not yield differences in SES levels for age groups. Another possible reason could be if the sample were somehow atypical. For instance, if the sample did not show the expected age effects on cognitive abilities (i.e., language, EF, Total ToM). Again, our analysis showed that the participants in the current study displayed differential performance by age in the expected direction for all cognitive measures. A final possibility for a confound could involve the fixed order in which the tasks were administered. Previous research has either compared several fixed orders or used a single fixed order. The earlier studies by Wellman and colleagues (Wellman & Liu, 2004; Wellman et al., 2006) have found no task order effect, which latter studies have used as a rationale for using a fixed order that starts with easier items and ends with the most complicated. The current study followed the same logic and presented children with a fixed order that increased in complexity, as documented by previous work. If children were uniformly inclined to do better or worse in the first vs. the last tasks, the order of tasks would act as a confound. However, different age groups found different tasks to be more or less challenging, although all tasks appeared in a fixed order for all ages.

Given that potential confounds do not seem to explain the current findings, it would be worthwhile to consider the potential scope of the results. First and foremost, we do not have information regarding whether the observed multiplicity of patterns by EF is exclusive to Turkish children or whether a similar multiplicity would be observed in different cultures. The current literature investigating developmental sequences in different cultures has administered Guttman analyses on age-pooled samples. This practice yields the most common sequence in a specific culture but does not have the power to tell us whether naturally occurring subsamples that are meaningfully different from one another (in our case by EF) coexist. To illustrate, in the current study, the Guttman analysis of the whole sample showed the most prevalent sequence to be the Chinese/Iranian pattern. However, when children were recategorized based on their EF ability, three patterns emerged for the different EF groups. A Guttman analysis by age mirrored these results. It is possible that samples classified as having either the Chinese/Iranian or the American/Australian pattern might include subsamples that show different patterns. Future work that either reanalyzes archival data or collects new data simultaneously in different cultures and uses such detailed analysis would yield insight into this question.

4.2. What if the multiplicity of sequences is the norm across cultures?

It would be prudent to consider explanations that do not rely on culture-specific factors if the multiplicity of patterns were cross-culturally the norm rather than the exception. Especially if the reported relations between sequences and EF were also replicated, it may be pertinent to question a foundational assumption of Wellman's constructivist framework for ToM development. Namely, is false belief understanding a conceptual accomplishment that builds on simpler prerequisite achievements? Alternatively, does EF act as a 'performance' variable that masks children's core ToM 'competence'? While the current study cannot resolve debates about the merit of questioning the foundational assumptions of constructivism, it certainly does not indicate a "competence-performance" based explanation to be true. Across the three patterns in our data, the DD, DB, and KA tasks were always passed before the false belief task (i.e., CFB). In other words, there is no evidence for eliminating the possibility that success on DD, DB, and KA tasks constitutes a prerequisite to the ability to attribute false beliefs and predict others' false-belief-driven behaviors (i.e., pass the CFB task). Hence, the weight-bearing beam in this hierarchy remains intact.

These results do not eliminate the possibility that children's FB understanding simultaneously relies on children's EF abilities and the conceptual development of their understanding of the mind. The fact that DD, DB, and KA show different patterns of success but are always passed before the false belief task (i.e., CFB) is also consistent with findings that show an Asian advantage in EF (e.g., Lan et al., 2011), but not in FB understanding (Sabbagh et al., 2006). In our study with Turkish children, shifts in the order with which the KA task was passed depended on children's EF ability. Children with high EF ability passed KA first, those with moderate EF passed it as the second task, and those with low EF passed it as the third task. However, the CFB task was passed as the fourth task for all children, regardless of their EF ability. This implies that success on the false belief task does not entirely depend on EF abilities. Since children passed these three tasks, of which only one (KA) was correlated with EF before they could pass the false belief task. Future cross-cultural research that includes the ToM scale and a battery of EF tasks that tap into EF components (inhibition, cognitive flexibility, and working memory) would answer this question more thoroughly.

While our data do not provide evidence for the idea that EF abilities could be solely responsible for children's ToM, especially false belief understanding as assessed by the CFB task, it does warrant a deeper consideration of the individual tasks and their cognitive demands. The individual tasks in the ToM scale may not be as comparable in cognitive burden as assumed. Wellman and Liu (2004) sought to control for secondary demands (length, number of stimuli), but it is impossible to remove task- and ability-specific cognitive requirements entirely. On the surface, all tasks seem to require some amount of EF. The DD and DB tasks may require children to entertain diverse attitudes and answer according to the attitude that does not belong to self. The KA and CFB tasks require children to inhibit their knowledge of reality to consider the world from a fictional character's perspective. The results of the current study provide some evidence for KA and CFB drawing upon similar cognitive resources. This is evidenced by the fact that they are correlated with each other and with EF when the effects of age are controlled. These findings are also in line with Doenyas et al. (2018). However, DD and DB correlate only with one another but not with EF. This result suggests that even if DD and DB are conceptual prerequisites to false belief understanding, they may not entirely share the same additional cognitive resources with KA and CFB.

Another important point for consideration is whether children pass or fail tasks with the kind of reasoning we researchers have in mind. It is developmentally reasonable to consider the possibility that children may succeed on some tasks with more basic and non-epistemic reasoning. Specifically, the tasks that include an understanding of diversity (i.e., DD and DB) could be passed without consideration of the assumed plurality of perspectives. That is, simple associationistic thinking, where children link the character in the scenario with the choice that is different from their own (desire or belief), would yield correct answers. In the current study, children's performance on the DD task (i.e., 81 %) was comparable to results from other studies with Turkish middle-class (82 % in Selcuk et al., 2018), Iranian (86 % in Shahaeian et al., 2011), and Chinese children (89 % in Wellman et al., 2006). It is also worth noting that in contrast to the other tasks in the scale, there was no age difference for DD or DB performance. Given that previous studies have not reported task performance by age (for an exception, see Sundqvist et al., 2018), it is impossible to know whether such stagnation is covertly present in the literature. The developmental stagnation present in the current data may not constitute a major problem for the DD task, where the results are comparable to other work. However, there seems to be a striking difference between the percentage of children who passed the DB task (57 %) in comparison to the reported percentages of children from Australia (85 % in Peterson et al., 2005), the US (84 % in Wellman & Liu, 2004), but also China (71 % in Wellman et al., 2006).

We argue that careful consideration should be given to the converse explanation: just as children can succeed in tasks through simpler forms of reasoning, they may fail due to epistemic reasoning that is culturally dependent. Examples of such 'regression' or 'stagnation' are evident for true vs. false belief task performance (Schidelko et al., 2022) and the over-imitation literature (e.g., Over & Carpenter, 2013). We will discuss this latter point in the next section, where we consider explanations for culture-specific variation.

4.3. *What if the observed variation is culture-specific?*

Turkish has a highly grammaticalized evidential component in which speakers encode the source of their knowledge through verbal affixes and particles (Aksu Koç, 2009). In the spirit of Dan Slobin's (1996) "thinking for speaking" framework, it is plausible to think that evidential features shape children's attention to particular aspects of speech. It is possible that evidentials designed to signal source information (e.g., hearsay vs. direct perception) may also inadvertently indicate reliability and certainty (e.g., hearsay being less reliable than direct perception). However, in light of previous studies which have shown a Turkish advantage in source monitoring (e.g., Aksu Koç, 2009; Aydin & Ceci, 2009; Lucas et al., 2013), the ascent of the KA task in task order by age, may also signal Turkish children's attention to the source of knowledge (i.e., seeing leads to knowing).

While analyzing culture-specific success profiles, it is pertinent to question not only the order with which the DB task is passed but also Turkish children's lower rate of success on the DB task (57 %), as compared to data that has been reported for American (84 %; Wellman & Liu, 2004) and Australian (85 %; Peterson et al., 2005) children. American and Australian children's early success in the DB task has been explained by virtue of their culture. Specifically, Wellman (2014) has conjectured that "individualistic, independent cultures where children are encouraged to think for themselves, to form and assert opinions freely, and to listen to others' varied views without privileging the traditional wisdom of elders over the creative new ideas of the young" (pp. 99–100) underlies American and Australian children's earlier success in the DB task. In addition, Wellman (2014) highlights the importance of "learning" and "knowing" in Chinese culture such that it may partly explain Chinese children's earlier success on the KA task. Interestingly, in the DB task, the Turkish children in our sample performed poorer than the Chinese children with whom they share the overall ToM scale sequence (71 % in both Duh et al., 2016; Wellman et al., 2006 compared to 57 % here). Further, Turkish children's performance on the DB task does not show significant change with age (i.e., 52 %, 57 %, 63 % for ages 3 to 5). Is it plausible that Turkish children do not make developmental progress in their appreciation of the diversity of beliefs despite showing expected improvement both in their false belief understanding and general cognitive abilities (e.g., language)? While this is possible, an alternative consideration would be to think carefully about the cross-cultural validity of the task.

Previously we mentioned how children could "succeed" on the DB task via simpler, associationist forms of reasoning. We believe it is also possible for children to fail while engaging in culture-specific epistemic reasoning. In the DB task, neither the fictional character nor the child knows with any certainty where the cat is hiding. Since there is a set of finite possibilities (i.e., under the car, in the bushes) and one is true, but no one knows the correct answer, the DB task can be construed as a belief ignorance task (for a different yet equally plausible conceptualization of this task (see Westra & Carruthers, 2017). In this task, children are asked, "Where do you *think* the cat is?" In essence, we expect children to *guess* in response to a question that uses the verb "*think*". While this may not be an uncommon practice in English (Zhang, 2014), Turkish language learners may be less frequently exposed to the use of "think" when speakers are less than sure or have no reason to be inclined to choose among viable options. Such a proposal requires future empirical work. However, we have some preliminary evidence from adult native speakers of Turkish that may provide a starting point. When 150 native adult Turkish speakers were presented with the DB task and were queried about the best way to ask the target question to preschoolers, among the two options (i.e., think, and guess), they opted for “guess” (75 %) rather than “think” (25 %) (Haskaraca & Ilgaz, 2021). Thus, the lack of developmental difference in DB performance across the age groups in the current study (i.e., 52 %, 57 %, 63 %) may not stem from cultural attitudes to diversity but from the way in which the task utilizes a mental state lexicon in ways that are culturally less prevalent.

The extant literature assumes that children across cultures understand the ToM tasks similarly if we use appropriate translations of the tasks. While cultures (and their concordant language systems) explicitly accommodate talk about mental states via mental state lexicon, the syntactic and pragmatic tools they use in conjunction may shape and nuance the meaning of seemingly perfect translations of the same words. Work with 3-year-old Japanese speakers, compared to German-speaking children, has shown that Japanese children benefit from the use of evidential syntax present in their language when making meaning of false belief statements (Matsui et al., 2009).

Future research should test whether Turkish-speaking children may be interpreting the question (i.e., “Where do you think the cat is”) to be asking for an informed opinion. During testing sessions, we frequently observed children look very closely at the picture stimuli and try to find clues as to where the cat might be. In reality, the pictures do not provide any clues. However, some children were observed talking to themselves while scrutinizing the pictures. We often heard children report evidence they imagined (“I can see the tail of the cat, it must be under the car/in the bushes”). It is plausible to think that committing to a version of reality through answering a question that asks for their “thoughts” when they have no evidence to form one, may confuse children whose language has a strongly grammaticalized component of evidentiality and uses “think/thought” for informed opinions. Future work that manipulates the pragmatic context along with mental state lexicon and matching grammar may begin to support this possibility.

The point about how children make meaning of the question may not be a trivial detail about what word to use in a task. Instead, it may suggest a substantive proposal: that there exists culture-specific ways of thinking and talking about epistemic concepts. This proposal is not new (Koenig, 2002; Lillard, 1998), but it has not yet been meaningfully incorporated into mainstream cognitive development research. The literature investigating how adults, teenagers, and children from different cultures *conceptualize* epistemological states (e.g., Hofer, 2008) has remained circumscribed to the pedagogical sciences. In the cognitive development literature, some of the early work on children’s understanding of different states of knowing could provide a starting point for future studies (e.g., Macnamara et al., 1976; Misciones et al., 1978; Moore et al., 1989, 1990; Schwanenflugel et al., 1996). Accordingly, we propose that future cross-cultural work on ToM that investigates conceptualizations of epistemic states in child and adult samples would provide a starting point for understanding inter-cultural differences in ToM development.

A second reason for culture-specific variance could be due to a variety of cultural and demographic variables that shape children’s social experiences. A non-exhaustive list includes parental education levels, parenting styles, parental cultural attitudes, and the length and type of preschool program children attend. While the current study had information on the number of siblings, information about their birth order and age differences were not asked. Recent work (Leblanc et al., 2017) shows that, for toddlers, having an older sibling may support theory of mind development but having a younger sibling may not. Given that only-children were found to perform as well as children with older siblings, the story is bound to be more complicated.

Another important factor that may shape characteristics of parent-child interaction and consequently affect ToM development is parents’ epistemological perspectives. Sociological and anthropological work can serve as a starting point for such an endeavor (Belenky et al., 1986). Some more recent work from psychology has sought to explore the possible relations between mothers’ epistemological perspectives, their behavior and communication with their children (e.g., communication strategies, mental state talk, interpretive talk), and child outcomes (e.g., interpretive ToM, vocabulary) (Bond & Burns, 2006; Hutchins et al., 2009; Tafreshi & Racine, 2016). Research that looks at parental epistemological perspectives in relation to sequence differences and children’s cognitive abilities (e.g., EF, language) may hold an important key to understanding intra- and inter-cultural differences in ToM development.

4.4. *Could there be room for both culture-specific and universal explanations?*

The current study showed that Turkish children’s performance on the KA task was related to their EF ability. However, their performance on the DB task was resistant to change. The current study does not have the power to explain the reasons behind these findings. However, plausible possibilities were presented for both findings. The results, in broad strokes, leave the universality argument intact. Regardless of age or EF ability, all children passed the DD, DB, and the KA task before they passed the false belief task (i.e., CFB). In this sense, we could talk of a universal sequence. However, there is also room for culture-specific explanations. So far, the variability in sequence has been taken to imply large-scale differences in cultural attitudes (e.g., collectivism vs. individualism), but empirical support for such conjecture has fallen short. In addition to work that includes measures of individualism/collectivism, work that studies the basic psychometric properties of the scale across cultures would add to our understanding of the socio-cognitive developments that precede and follow false belief understanding. Currently, the scale is simultaneously used as conceptual grounds and evidence of theory of mind development. Both the theory behind the scale and the measure itself would benefit from developing reliability and validity studies. This would include consideration of the cross-cultural validity for each task in the scale. In other words, despite literal translations, the task may not actually be measuring the same ability across cultures. If this is the case, we propose that such considerations have more than mere methodological value. We should systematically and cross-culturally study the developmental makeup of epistemic concepts and their manifestation through language. Such a practice would be able to better inform us whether the relations between cognitive abilities and sequence differences are universal or culture-specific.

4.5. Conclusion

Extant conceptions of belief have been assumed to be similar across cultures. This is evidenced by the common practice of using direct translations of ToM scale tasks with minimal adjustments (culturally relevant stimuli). Extant conceptions of ToM development have also assumed similarity across ages. This is evidenced by studies generally taking preschoolers as a single age group in order to find the most prevalent pattern (see for an exception, Sundqvist et al., 2018). Moreover, performance differences have been generally attributed to a variable (i.e., cultural attitudes) that either was not measured (e.g., Wellman et al., 2006); when measured, did not reveal the expected differences (e.g., Kabha & Berger, 2020); or was measured with a proxy variable that showed the opposite of the expected results (Selcuk et al., 2018).

The current study contributes to the literature by showing sequence differences by age and EF abilities in a Turkish sample. The prevalent explanation that appeals to the distinction between individualism vs. collectivism would seem to have difficulty explaining the current findings. If there are no confounding socio-demographic differences among the current sample, why would 3-year-olds show one pattern and 4-year-olds show another pattern? Furthermore, why would 5-year-olds show yet a different pattern? The current study does not have the power to explain these differences, but these findings provide the impetus for various types of future studies that are yet to be undertaken. By situating false belief understanding in the context of developing epistemic abilities Wellman and colleagues have opened the door for whether children's understanding of the mind follows similar or divergent paths across cultures. A comprehensive and accurate answer will likely be nuanced and will involve longitudinal studies that investigate the interactive effects of several aspects of cultural reality and cognitive abilities.

Declaration of Interest

None.

References

Aksu Koç, A. (2009). Evidentials: An interface between linguistic and conceptual development. In J. Guo, E. Lieven, N. Budwig, S. Ervin-Tripp, K. Nakamura, & Ş. Özçalışkan (Eds.), *Crosslinguistic approaches to the psychology of language: Research in the tradition of Dan Isaac Slobin* (pp. 531–542). Psychology Press.

Aydin, Ç., & Ceci, S. J. (2009). Evidentiality and suggestibility: A new research venue. *New Directions for Child and Adolescent Development, 125*, 79–93. https://doi.org/10.1002/cd.251

Belenky, M. F., Clincher, B. M., Goldberger, N. R., & Tarule, J. M. (1986). *Women's ways of knowing: The development of self, voice and mind*. Basic Books.

Berument, S. K., & Güven, A. G. (2013). Turkish expressive and receptive language test: I. Standardization, reliability and validity study of the receptive vocabulary sub-scale. *Türk Psikiyatri Dergisi, 24*(3), 192–201.

Bond, L. A., & Burns, C. E. (2006). Mothers' beliefs about knowledge, child development, and parenting strategies: Expanding the goals of parenting programs. *Journal of Primary Prevention, 27*(6), 555–571. https://doi.org/10.1007/s10935-006-0061-9

Bowerman, B. L., & O'Connell, R. T. (1990). *Linear statistical models: An applied approach*. Duxbury.

Callaghan, T., Rochat, P., Lillard, A., Claux, M. L., Odden, H., Itakura, S., … Singh, S. (2005). Synchrony in the onset of mental-state reasoning: Evidence from five cultures. *Psychological Science, 16*(5), 378–384. https://doi.org/10.1111/j.0956-7976.2005.01544.x

Devine, R. T., & Hughes, C. (2016). Measuring theory of mind across middle childhood: Reliability and validity of the silent films and strange stories tasks. *Journal of Experimental Child Psychology, 149*, 23–40. https://doi.org/10.1016/j.jecp.2015.07.011

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

Dörrenberg, S., Rakoczy, H., & Liszkowski, U. (2018). How (not) to measure infant theory of mind: Testing the replicability and validity of four non-verbal measures. *Cognitive Development, 46*, 12–30. https://doi.org/10.1016/j.cogdev.2018.01.001

Duh, S., Paik, J. H., Miller, P. H., Gluck, S. C., Li, H., & Himelfarb, I. (2016). Theory of mind and executive function in Chinese preschool children. *Developmental Psychology, 52*(4), 582–591. https://doi.org/10.1037/a0040068

Dunn, L. M. & Dunn, D. M. (2007). *PPVT-4: Peabody picture vocabulary test*. Pearson Assessments.

Etel, E., & Yagmurlu, B. (2015). Social competence, theory of mind, and executive function in institution-reared Turkish children. *International Journal of Behavioral Development, 39*(6), 519–529. https://doi.org/10.1177/0165025414556095

Gopnik, A., & Wellman, H. M. (1992). Why the child's theory of mind really is a theory. *Mind & Language, 7*(1–2), 145–171. https://doi.org/10.1111/j.1468-0017.1992.tb00202.x

Green, B. F. (1956). A method of scalogram analysis using summary statistics. *Psychometrika, 21*(1), 79–88. https://doi.org/10.1007/BF02289088

Haskaraca, F., & Ilgaz, H. (2021, January 4-8). *Turkish speakers' conceptualization of belief-related words and its implications for Theory of Mind tasks* [Poster presentation]. Budapest CEU Conference on Cognitive Development.

Henning, A., Spinath, F. M., & Aschersleben, G. (2011). The link between preschoolers' executive function and theory of mind and the role of epistemic states. *Journal of Experimental Child Psychology, 108*(3), 513–531. https://doi.org/10.1016/j.jecp.2010.10.006

Hofer, B. (2008). Personal epistemology and culture. In M. S. Khine (Ed.), *Knowing, knowledge and beliefs: Epistemological studies across diverse cultures* (pp. 3–22). Springer Science.

Hutchins, T., Bond, L., Silliman, E., & Bryant, J. (2009). Maternal epistemological perspectives and variations in mental state talk. *Journal of Speech, Language, and Hearing Research, 52*, 61–80. https://doi.org/10.1044/1092-4388(2008/07-0161)

Kabha, L., & Berger, A. (2020). The sequence of acquisition for theory of mind concepts: The combined effect of both cultural and environmental factors. *Cognitive Development, 54*, Article 100852. https://doi.org/10.1016/j.cogdev.2020.100852

Kagitcibasi, C., & Ataca, B. (2005). Value of children and family change: A three-decade portrait from Turkey. *Applied Psychology, 54*(3), 317–337. https://doi.org/10.1111/j.1464-0597.2005.00213.x

Koenig, M. A. (2002). Children's understanding of belief as a normative concept. *New Ideas in Psychology, 20*(2–3), 107–130. https://doi.org/10.1016/S0732-118X(02)00004-1

Kristen, S., Thoermer, C., Hofer, T., Aschersleben, G., & Sodian, B. (2006). Validation of the "theory of mind" scale. *Zeitschrift fur Entwicklungspsychologie und Paedagogische Psychologie, 38*(4), 190–199.

Lan, X., Legare, C. H., Ponitz, C. C., Li, S., & Morrison, F. J. (2011). Investigating the links between the subcomponents of executive function and academic achievement: A cross-cultural analysis of Chinese and American preschoolers. *Journal of Experimental Child Psychology, 108*(3), 677–692. https://doi.org/10.1016/j.jecp.2010.11.001

Leblanc, É., Bernier, A., & Howe, N. (2017). The more the merrier? sibling composition and early manifestations of theory of mind in toddlers. *Journal of Cognition and Development, 18*(3), 375–391. https://doi.org/10.1080/15248372.2017.1327438

Lewis, C., Freeman, N. H., Kyriakidou, C., Maridaki-Kassotaki, K., & Berridge, D. M. (1996). Social influences on false belief access: specific sibling influences or general apprenticeship? *Child Development, 67*(6), 2930–2947. https://doi.org/10.1111/j.1467-8624.1996.tb01896.x

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin, 123*(1), 3–32. https://doi.org/10.1037/0033-2909.123.1.3

Lucas, A. J., Lewis, C., Pala, F. C., Wong, K., & Berridge, D. (2013). Social-cognitive processes in preschoolers' selective trust: three cultures compared. *Developmental Psychology, 49*(3), 579–590. https://doi.org/10.1037/a0029864

Macnamara, J., Baker, E., & Olson, C. L. (1976). Four-year-olds' understanding of pretend, forget, and know: Evidence for propositional operations. *Child Development, 47*(1), 62–70. https://doi.org/10.2307/1128283

Matsui, T., Rakoczy, H., Miura, Y., & Tomasello, M. (2009). Understanding of speaker certainty and false-belief reasoning: A comparison of Japanese and German preschoolers. *Developmental Science, 12*, 602–613. https://doi.org/10.1111/j.1467-7687.2008.00812.x

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

Misciones, J. L., Marvin, R. S., O'Brien, R. G., & Greenberg, M. T. (1978). A developmental study of preschool children's understanding of the words 'know' and 'guess'. *Child Development, 49*(4), 1107–1113. https://doi.org/10.2307/1128750

Moore, C., Bryant, D., & Furrow, D. (1989). Mental terms and the development of certainty. *Child Development, 60*(1), 167–171. https://doi.org/10.2307/1131082

Moore, C., Pure, K., & Furrow, D. (1990). Children's understanding of the modal expression of speaker certainty and uncertainty and its relation to the development of a representational theory of mind. *Child Development, 61*(1), 722–730. https://doi.org/10.2307/1131082

Moses, L. J., & Tahiroglu, D. (2010). Clarifying the relation between executive function and children's theories of mind. In J. Carpendale, G. Iarocci, U. Muller, B. Sokol, & A. Young (Eds.), *Self- and social regulation: Exploring the relations between social interaction, social cognition, and the development of executive functions* (pp. 218–231). Oxford University Press.

Myers, R. (1990). *Classical and modern regression with applications* (2nd ed.). Duxbury.

Over, H., & Carpenter, M. (2013). The social side of imitation. *Child Development Perspectives, 7*(1), 6–11.

Özdikmenli-Demir, G., & Sayıl, M. (2009). Individualism-collectivism and conceptualizations of interpersonal relationships among Turkish children and their mothers. *Journal of Social and Personal Relationships, 26*(4), 371–387. https://doi.org/10.1177/0265407509350557

Perner, J., Ruffman, T., & Leekam, S. R. (1994). Theory of mind is contagious: You catch it from your sibs. *Child Development, 65*(4), 1228–1238. https://doi.org/10.1111/j.1467-8624.1994.tb00814.x

Peterson, C., & Slaughter, V. (2003). Opening windows into the mind: Mothers' preferences for mental state explanations and children's theory of mind. *Cognitive Development, 18*(3), 399–429. https://doi.org/10.1016/S0885-2014(03)00041-8

Peterson, C. C., Wellman, H. M., & Liu, D. (2005). Steps in theory-of-mind development for children with deafness or autism. *Child Development, 76*(2), 502–517. https://doi.org/10.1111/j.1467-8624.2005.00859.x

Sabbagh, M. A., Xu, F., Carlson, S. M., Moses, L. J., & Lee, K. (2006). The development of executive functioning and theory of mind: A comparison of Chinese and US preschoolers. *Psychological Science, 17*(1), 74–81. https://doi.org/10.1111/j.1467-9280.2005.01667.x

Schidelko, L. P., Huemer, M., Schröder, L. M., Lueb, A. S., Perner, J., & Rakoczy, H. (2022). Why do children who solve false belief tasks begin to find true belief control tasks difficult? A test of pragmatic performance factors in theory of mind tasks. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.797246

Schwanenflugel, P. J., Fabricius, W. V., & Noyes, C. R. (1996). Developing organization of mental verbs: Evidence for the development of a constructivist theory of mind in middle childhood. *Cognitive Development, 11*(2), 265–294. https://doi.org/10.1016/S0885-2014(96)90005-2

Scott, R. M., & Baillargeon, R. (2017). Early false-belief understanding. *Trends in Cognitive Sciences, 21*(4), 237–249. https://doi.org/10.1016/j.tics.2017.01.012

Selcuk, B., Brink, K. A., Ekerim, M., & Wellman, H. M. (2018). Sequence of theory-of-mind acquisition in Turkish children from diverse social backgrounds. *Infant and Child Development, 27*(4), Article e2098. https://doi.org/10.1002/icd.2098

Shahaeian, A. (2015). Sibling, family, and social influences on children's theory of mind understanding: New evidence from diverse intracultural samples. *Journal of Cross-Cultural Psychology, 46*(6), 805–820. https://doi.org/10.1177/0022022115583897

Shahaeian, A., Peterson, C. C., Slaughter, V., & Wellman, H. M. (2011). Culture and the sequence of steps in theory of mind development. *Developmental Psychology, 47*(5), 1239–1247. https://doi.org/10.1037/a0023899

Shahaeian, A., Nielsen, M., Peterson, C. C., Aboutalebi, M., & Slaughter, V. (2014). Knowledge and belief understanding among Iranian and Australian preschool children. *Journal of Cross-Cultural Psychology, 45*(10), 1643–1654. https://doi.org/10.1177/0022022114548484

Slobin, D. I. (1996). From "thought and language" to "thinking for speaking". In J. J. Gumperz, & S. C. Levinson (Eds.), *Rethinking linguistic relativity* (pp. 70–96). Cambridge, England: Cambridge University Press.

Sundqvist, A., Holmer, E., Koch, F., & Heimann, M. (2018). Developing theory of mind abilities in Swedish pre-schoolers. *Infant and Child Development, 27*(4), 747–760. https://doi.org/10.1002/icd.2090

Tafreshi, D., & Racine, T. P. (2016). Children's interpretive theory of mind: The role of mothers' personal epistemologies and mother-child talk about interpretation. *Cognitive Development, 39*, 57–70. https://doi.org/10.1016/j.cogdev.2016.04.003

TR Ministry of Labor, Social Services, and Family. (2018). *Net minimum wages by years*. ⟨https://birim.ailevecalisma.gov.tr/En/Contents/Istatistikler/AsgariUcret⟩.

Wellman, H. (1990). *The child's theory of mind*. MIT Press.

Wellman, H. (2014). *Making minds: How theory of mind develops*. Oxford University Press.

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*(3), 655–684. https://doi.org/10.1111/1467-8624.00304

Wellman, H. M., Fang, F., Liu, D., Zhu, L., & Liu, G. (2006). Scaling of theory-of-mind understandings in Chinese children. *Psychological Science, 17*(12), 1075–1081. https://doi.org/10.1111/j.1467-9280.2006.01830.x

Westra, E., & Carruthers, P. (2017). Pragmatic development explains the Theory-of-Mind Scale. *Cognition, 158*, 165–176. https://doi.org/10.1016/j.cognition.2016.10.021

Williams, K. T. (2007). *Expressive vocabulary test-second edition (EVT-2)*. Pearson Assessments.

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

Yagmurlu, B., Berument, S. K., & Celimli, S. (2005). The role of institution and home contexts in theory of mind development. *Journal of Applied Developmental Psychology, 26*(5), 521–537. https://doi.org/10.1016/j.appdev.2005.06.004

Zelazo, P. D. (2006). The dimensional change card sort (DCCS): A method of assessing executive function in children. *Nature Protocols, 1*, 297–301. https://doi.org/10.1038/nprot.2006.46

Zhang, G. (2014). The elasticity of I think: Stretching its pragmatic functions. *Intercultural Pragmatics, 11*(2), 225–257. https://doi.org/10.1515/ip-2014-0010