# Language and false belief: Evidence for general, not specific, effects in Cantonese-speaking preschoolers

**Article** *in* Developmental Psychology · April 2007

DOI: 10.1037/0012-1649.43.2.318 · Source: PubMed



**3 authors:**

Twila Tardif
University of Michigan


Wing Chee So
The Chinese University of Hong Kong


Niko Kaciroti
University of Michigan

Developmental Psychology
2007, Vol. 43, No. 2, 318–340

0012-1649/07/$12.00 DOI: 10.1037/0012-1649.43.2.318

# Language and False Belief: Evidence for General, Not Specific, Effects in Cantonese-Speaking Preschoolers

Twila Tardif
University of Michigan

Catherine Wing-Chee So
University of Chicago

Niko Kaciroti
University of Michigan

Two studies were conducted with Cantonese-speaking preschoolers examining J. de Villiers's (1995) hypothesis that syntactic complements play a unique role in the acquisition of false belief (FB). In Study 1, the authors found a positive correlation between FB and syntactic complements in 72 four- to six-year-old Cantonese-speaking preschoolers. Study 2 followed 72 three- to five-year-old Cantonese-speaking children who initially failed an FB screening task and were then tested on general language abilities, short-term memory, inhibition, nonverbal IQ, and on FB and complement tasks. Once age and initial FB understanding were controlled for in both multiple regression and hierarchical linear modeling analyses, complements no longer uniquely predicted FB. Instead, individual differences in general language abilities and short-term memory contributed to the variation in both complements and FB.

*Keywords:* theory of mind, language, complements, Chinese, hierarchical linear modeling

Between the ages of 3 and 6, young children begin to develop an understanding of false belief (FB). Specifically, they go from being able to understand simple desires and distinctions between one's own and another person's desires well before their 2nd birthday (Bartsch & Wellman, 1995) to understanding beliefs and, more slowly, the distinction between one's own and another person's beliefs sometime before the age of 6 (Wellman & Liu, 2004). This general sequence, with some variations in timing, is true in Western cultures, but it is also true in all of the non-Western cultures (e.g., Chinese, Japanese, Korean, Mofu, Quechuan) that have been studied so far (Flavell, 2004; Liu, Wellman, Tardif, & Sabbagh, 2006; Vinden, 1996, 1999, 2001, 2002; Wellman, Cross, & Watson, 2001). Although there is much agreement on the fact of this development, explanations for it are fraught with controversy. Some explanations focus on children's developing cognitions about the world (Gopnik & Wellman, 1994; Perner, 1991; Wellman & Gelman, 1992), whereas others focus on the role of other developments or experiences in jump-starting or impeding this process (e.g., Astington, 2000; Carlson, Moses, & Breton, 2002; Hughes, Dunn, & White, 1998; Lillard, 1998; Olson, 1988; Vinden, 1996). Children's developing language abilities, in particular, have been identified as one factor that researchers have identified as having a direct influence on FB performance and theory of mind (ToM) development more generally. Nonetheless, researchers who have examined the effects of language on FB performance vary greatly in whether they find this to be a general effect of language or whether there are more specific linguistic factors that influence the understanding of FB.

Twila Tardif and Niko Kaciroti, Department of Psychology, University of Michigan; Catherine Wing-Chee So, Department of Psychology, University of Chicago.

This research was made possible by a grant from the Research Grants Council of the Hong Kong Special Administrative Region (Project #CUHK4116/99H) and by funding from the Center for Human Growth and Development at the University of Michigan.

We thank Jill de Villiers for providing pictures and sentences from her false complements task as well as the children, their parents, and the preschools who participated. In addition, we thank Angel Chan, Kitty Yau Fong Fung, and Apple Au-Yeung, who helped with the data collection, and Kawai Leung and Joey Fung, who helped with manuscript preparation.

Correspondence concerning this article should be addressed to Twila Tardif, The Center for Human Growth and Development, University of Michigan, 300 North Ingalls, 10th Floor, Ann Arbor, MI 48109-0406. E-mail: twila@umich.edu

## General Effects for the Supporting Role of Language in FB Understanding

The evidence for more general effects of language on FB understanding suggests both that an adequate amount of exposure to talk and, more specifically, talk about mental states, must occur in the child's environment in order to acquire FB. For example, deaf children of hearing parents are greatly delayed in FB understanding in relation to hearing children and also in relation to deaf children of deaf (signing) parents (Gale, de Villiers, de Villiers, & Pyers, 1996; Peterson & Siegal, 1995). Once deaf children start to sign and become exposed to more complex language, including mental state language, through signing, their FB understanding catches up accordingly. It is important to note that this relationship holds even when the children are tested on nonverbal tests of FB (Gale et al., 1996). Autistic children, who are delayed in their linguistic development, also show severe delays in FB performance (Tager-Flusberg, 1993), and those who do pass FB tasks tend to have higher verbal abilities than those who do not pass these tasks (Happé, 1995). Finally, there is also a clear relationship between performance on FB tests and standardized measures of language ability in normally developing hearing children (Cutting & Dunn, 1999; Hughes & Dunn, 1997; Jenkins & Astington, 1996). Thus for both hearing and deaf individuals, not only is language generally facilitative of FB understanding, but a child's level of linguistic sophistication bears a direct relationship to the ability to understand FB.

## Effects of Specific Lexical and Syntactic Acquisitions on FB Understanding

A more controversial claim is whether the acquisition of specific lexical items or specific syntactic structures underlies this general linguistic influence on children's ability to understand FB. Olson (1988), for instance, argued that children's acquisition of mental state terms such as *know* and *think* is critical for developing a theory of mind that would allow them to pass FB tasks. Although not a direct test of this hypothesis, Moore, Pure, and Furrow (1990, Experiment 2) found that, for English-speaking children, there was indeed a relation between children's comprehension of *know* versus *think* and their ability to understand FB.

### *Lexical Effects*

Other researchers have examined how the mental state language of particular languages might be related to children's theory of mind development. English, for example, has several ways of talking about the mental states of thinking and knowing, including *think*, *believe*, *guess*, *know*, and so on, each of which involves differences in the thinker's certainty and other aspects of their mental states (Moore, Bryant, & Furrow, 1989). With respect to FB, in English, belief terms such as *think* and *believe* are neutral with respect to whether the person's belief is true or false. If told that "Joe *thinks* the ball is red," it may be red or blue. Some languages, including Chinese languages, include terms to designate beliefs that are decidedly false. The appearance of such terms has several intriguing implications. At the extreme, perhaps children learning such a language never evidence difficulty with FB because their language clearly marks this distinction. Alternatively, understanding FB may be enhanced but more modestly. Or the presence of explicit FB terms may provide no enhancement whatsoever. Three separate cross-linguistic studies involving Mandarin, Cantonese, English, Puerto Rican Spanish, and Turkish provide converging evidence for a very limited facilitation of FB performance when words that explicitly mark FB are used in the test question. In all three studies, the use of the words that explicitly mark FB for languages which have them appears to help children demonstrate FB understanding but only when they are used in the test question and only when that understanding is already emergent (Lee, Olson, & Torrance, 1999; Shatz, Diesendruck, & Martinez-Beck, 2003; Tardif, Wellman, & Cheung, 2004). In contrast, for children whose native language has multiple words marking different types of beliefs, use of a neutral verb appears to cause difficulty in an FB task. A critical comparison in all three of these studies, however, is that even when the FB verbs are used in the test questions, children speaking languages that mark FB explicitly are at no additional advantage when compared with children whose languages (e.g., English) have more neutral or polysemous verbs for belief.

Thus, findings on the use of specific lexical terms show that the effects of language on theory of mind development can go beyond the general effect of comparing exposure versus lack of exposure to mental state talk, demonstrated in studies of deaf children (e.g., Gale et al., 1996; Peterson & Siegal, 1995). However, the use of a specific lexical term for FB is only facilitative when the term is used in the question and when children are already at the cusp of understanding FB (Tardif et al., 2004).

### *Syntactic Effects*

Research on the role of children's syntactic abilities in facilitating and/or limiting FB understanding has also generated hypotheses about both the general effects of syntax as well as specific syntactic structures on the understanding of FB. In a longitudinal study examining the relationships between children's language abilities and theory of mind (including FB) understanding, Astington and Jenkins (1999) found a strong and unique influence of children's syntactic abilities. It is important to note that they found that language abilities predicted theory of mind and not the reverse, and that syntax, but not semantics, was the important contributor to this relationship.

At an even more specific level, de Villiers (1995; de Villiers & de Villiers, 2000) has proposed that the relationship between language and FB stems from a particular syntactic structure that is required for statements of FB—the finite complement structure. According to de Villiers, "the complex syntax that is used for describing mental events *makes possible* the representational changes that allow for understanding false beliefs" (1995, p. 1, emphasis added). As evidence for this, Roeper and de Villiers (1994) demonstrated that in an animated story in which Big Bird has a party to which he invited his friend Grover (but later forgets this fact) and to which he forgot to invite Bert, 4- and 5-year-old children were not able to distinguish between the following types of predicate complements.

1. Who did Big Bird forget to invite? (Bert)
2. Who did Big Bird forget that he invited? (Grover)

Instead, they answered the second type of question as if they understood it to be the first type of question—suggesting that although they might be able to understand the infinitive complement structure in Sentence 1, they were not yet able to understand the finite complements in Sentence 2. It is precisely this structure, however, that must be understood in talk about FBs. Thus, according to de Villiers, the semantic developments required for FB understanding cannot emerge until children are able to understand the syntactic structures in which such talk is embedded.

In English, a handful of studies have now examined the relationship between finite complements and FB directly by using both longitudinal and training study methods and found evidence that comprehension of complements does in fact appear to have an effect on FB understanding, although authors are divided as to how this relationship should be interpreted. De Villiers and Pyers (2002) conducted a longitudinal study of 28 native-English-speaking preschoolers whose FB and complement understanding (of both communication and mental state verbs) were tested at four time points together with spontaneous speech recordings at each time point. Overall, this study found that English-speaking children's FB understanding could be uniquely predicted by measures of syntax that included complements and by the separate measure of complement understanding (a memory for complements test) but not by measures of syntax that excluded complements. However, there are a number of limitations to this study as well. First, the small number of children ($n = 28$) may have limited the predictive power of other linguistic variables. Second, the regression analyses in this study did not control for age or initial FB performance. If these factors had been controlled, the unique variance attributed to complements would likely have been reduced (see, e.g., Astington & Jenkins, 1999; Hughes, 1998b; see the present Study 2).

Two more recent training studies, however, suggest that the link between complements and FB understanding may still be important, at least for English-speaking children. Specifically, Hale and Tager-Flusberg (2003) provided training for three groups of children in FB, sentence complements, or relative clauses, and they found, first, that children improved performance for the task in which they were trained. In other words, children trained on FB tasks performed better on FB tasks and children trained on relative clauses performed better on relative clause tasks after training. In addition, the sentence complements group showed a significant improvement in FB performance (their improvement in FB was equal to their improvement in complements, and their FB improvement was also equal to that of the group, which was specifically trained on FB), whereas neither the relative clause nor the FB group showed any transfer of training effects beyond the task in which they were trained. Thus, it appears that training children in comprehending and remembering complements can help them to pass FB tasks but not vice versa. Nonetheless, it is still not clear from this study whether it was the complement syntax per se that was responsible for the improvement in FB or whether it was simply discourse about the type of situation involved in both sentence complements and FBs—the need to coordinate across clauses and situations that have different truth values—that might have led to the improved performance.

In a different training study by Lohmann and Tomasello (2003), this hypothesis was examined directly, and between 24 and 30 children were assigned to each of five training conditions: full training about deceptive objects with mental verbs and sentence complements; full training about deceptive objects with communication verbs and sentence complements; discourse about deceptive objects without mental state verbs or complement constructions; simple commands to "Look!", highlighting the deceptive aspect of the objects (no language condition); or sentence complement training without reference to the deceptive aspect of the objects. As with previous studies (e.g., de Villiers & Pyers, 2002), Lohmann and Tomasello did not find a difference between the mental state versus communication verb conditions, but they did find important differences between the remaining conditions. Specifically, the no language group performed most poorly, even though the deceptive aspect of the objects was highlighted in this condition, and the full training groups performed the best. Moreover, the discourse only and complement only groups had almost identical scores and were intermediate between the no language and full training conditions. This suggests that although sentence complements can certainly facilitate FB understanding, they are not the only path toward that understanding—general discourse (that goes beyond "Look!" but does not include mental state verbs or finite complements) about an object's deceptive appearance can also help improve children's performance on an FB task. Nonetheless, when discourse about deceptive objects includes complements, performance is better still.

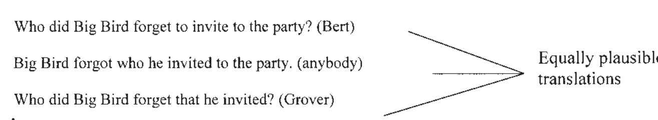
### *Complements and FB Across Languages*

Researchers who focus on languages (e.g., Chinese, German) for which distinctions in complement structures do not correspond with the belief–desire contrast have questioned this hypothesis more fundamentally for the simple reason that it does not account for the change in FB understanding that also occurs in children who speak these other languages (Perner, Sprung, & Zauner, 2003; Tardif & Wellman, 2000). As noted by Tardif and Wellman (2000), it is possible to translate both of the Big Bird sentences into Mandarin and Cantonese in ways that clarify the distinction (Sentences 3 and 4) or in ways that make it ambiguous (Sentence 5), but in neither case is the distinction purely a morphosyntactic one. Instead, children must learn to make distinctions based on their knowledge of the context as well as from variations in relative stress and word order.

3. Big Bird *m4-gei3-dak1 (zo2) ceng2 bin1go3*? (Bert)
Big Bird forget ASPECT invite who
Who (did) Big Bird forget (to) invite?

4. Big Bird *m4-gei3-dak1 ceng2 zo2 bin1go3*? (Grover)
Big Bird forget invite ASP who
Big Bird forgot (he) invite(d) who?

5. Big Bird *m4-gei3-dak1 zo2 ceng2 bin1go3 heoi3* party? (ambiguous)
Big Bird forget ASP invite who go party




Instead of the use of syntax, then, Mandarin and Cantonese mark FB understanding with the use of different lexical items (e.g., *yi3wei2/ji5wai4* vs. *xiang3/nam5*) to mark whether a belief is true or false and also to distinguish desires from beliefs ([xiang3]+yao4/[soeng5]+jiu4 vs. yi3wei2/ji5wai4). And, as mentioned above, when the appropriate FB term (*yi3wei2/ji5wai4*) is used, children perform much better than when the neutral term (*xiang3/nam5*) is used (Lee et al., 1999; Tardif et al., 2004). Moreover, in a recent study that examined the relationship between complement structures for *say* and *want*, which share the same complement syntax as *think* (*ji5wai4*), Cheung and colleagues (Cheung et al., 2004) found that although there was a positive correlation between complement and FB understanding, it was not specific to one or the other of these verbs (*say* is marked by finite complements in English, whereas *want* is expressed with nonfinite complements, but they are expressed with the same nonfinite syntax in Cantonese), and this relationship disappeared once general language abilities were considered in the same regression model.

German can also use a single syntactic form to express both desire and belief, but it does so by using finite complements for both types of sentences when a third person's mental states are the focus (Perner et al., 2003). Nonetheless, in Perner et al.'s (2003) study, even the youngest (3-year-old) German-speaking children were found to understand finite complement sentences for desires (*will . . . dass*), but only the older (5-year-old) children were able to understand the same types of sentences for thoughts and beliefs (*glaubt . . . dass*).

Although both the Chinese and the German studies suggest reasons to be wary of the link between complements and FB understanding, the Chinese studies did not directly examine the type of complements that de Villiers (1995) specifically emphasized in her hypothesis—false complements, particularly for verbs of communication and belief. Moreover, neither the Cheung et al. (2004) study of Cantonese-speaking nor the Perner et al. (2003) study of German-speaking children's development of finite complements and FBs examined this relationship longitudinally in relation to individual children's changes in understanding of these concepts.

Finally, none of these studies (in English, German, or Cantonese) examined whether other factors such as working memory or inhibitory abilities, both components of executive functioning, which also improve dramatically during the preschool years (Hughes, 1998a), might be related to both FB and complement understanding and perhaps offer a less language-specific explanation for the relationship. In a longitudinal study of children's executive functioning and FB performance, for instance, Hughes (1998b) found that children's working memory (measured by both traditional short-term memory tasks and second-order "working" memory tasks), planning, and inhibition abilities together accounted for 20% of the variation in FB performance, even after individual differences in linguistic abilities, initial FB performance, and age were controlled for. Thus, in order to more fully explore the relationship between language (both broadly and with respect to specific syntactic effects) and FB, one not only must look to other languages for which the relationship may not be so specific but also must consider the role of other underlying factors in mediating that relationship.

To that end, this article includes two studies that explore directly the role of complement structures in predicting FB understanding in Cantonese-speaking preschoolers. The first study is a cross-sectional one that seeks to discover whether there is a role for false complements in FB understanding in Cantonese, even though it is possible to produce both desire and belief statements with the same syntactic structures and without relying on sentence complements. In essence, our first study seeks to replicate Cheung et al.'s (2004) findings but with the specific types of FB and communication verb complements de Villiers (1995; de Villiers & de Villiers, 2000) predicted to be most relevant to her hypothesis. In the second study, having found a positive relationship between complements and FB, even in Cantonese, we explore this relationship longitudinally. Specifically, we trace the development of complement and FB understanding at four testing points over a 6-month period to try and determine the direction and possible reasons for this relationship. In addition, we also include a number of control variables (nonverbal IQ, inhibitory control, short-term memory, and a general language measure) to determine whether this relationship is mediated or even completely subsumed by these other factors that may be related to the development of both complement understanding and FB.

# Study 1

## *Method*

### *Participants*

Seventy-two native Cantonese-speaking preschoolers participated, with 24 children, half male and half female, at each of the three age groups. The 4-year-olds ranged from 4.0 to 4.12 (age: $M = 4.48$, $SD = 0.30$), the 5-year-olds ranged from 5.0 to 5.12 (age: $M = 5.40$, $SD = 0.34$), and the 6-year-olds ranged from 6.0 to 6.8 (age: $M = 6.18$, $SD = 0.15$), with birth date information provided by parents and age at the date of testing calculated from this information. All children were ethnic Chinese attending local kindergartens in a suburban area of Hong Kong. Although it was not possible to obtain further socioeconomic information from our sample, the kindergartens selected for the study were private, nonsubsidized kindergartens in moderate- to middle-income housing areas of Hong Kong. They were neither the best nor the worst in their category and were thus reflective of average to above-average kindergartens in the city.

### *Research Design*

We were interested in examining the relationship between sentence complements and FB across the preschool years. Thus, we considered age and gender to be our primary independent variables, with memory for complements and FB understanding as our primary dependent variables.

### *Materials and Scoring*

*Memory for complements.* We developed our complements task from de Villiers's (1995) memory for complements task in English in which children were asked about the contents of eight different sentences.[1] Most of the original sentences from de Villiers's (1995) study were used and simply translated into Cantonese with some changes to make them more relevant to local children. The translations were checked and revised by several native Cantonese-speaking and native English-speaking bilinguals until each sentence was clearly using Cantonese complement syntax and was parallel to the original English sentence. A full list of these sentences is provided in the Appendix.

[1] One criticism of both our studies and those done by de Villiers (de Villiers, 1995; de Villiers & de Villiers, 2000) is that the memory for complements task itself could involve FB and thus any relationship between the two tasks would be trivial. Although this is a possibility and is certainly the case when FBs are expressed in false complement statements, children given nonverbal versions of the task or test questions involving "look" show similar patterns of FB development (see Wellman et al., 2001). Thus, one can test FB understanding without the use of FB verbs or verb complement structures. Similarly, children often produce complements that are not false and can produce false complements without the involvement of belief (see de Villiers & de Villiers, 2000). Indeed, in our studies, as in those conducted by de Villiers (de Villiers, 1995; de Villiers & de Villiers, 2000), half of the memory for complements test questions involve communication verbs. De Villiers's argument, nonetheless, is that children must be able to understand and reproduce complement *structures* before they are able to achieve an understanding of FB.

Like de Villiers (1995), each sentence also had an accompanying pair of photographs such that the scenario in each set of photographs involved a character who made mistakes, told lies, or had an FB. Of the eight sets of photographs/sentences, half involved acts of thinking (e.g., The teacher *thought* the girl was reading, but really she was playing cards), and the other half involved acts of communication (e.g., The teacher *told* the girl there was a bug in her hair, but really it was a leaf). For each scenario, the experimenter presented the set of photographs while saying the sentence and pointing to the relevant photograph for each component of the sentence. Then, after she said the sentence and presented the photographs, the experimenter asked the child about either the "true" (reality/control) component of the sentence (i.e., What was she *doing*? *playing cards*; *gam2 go3 neoi4zai2 zou6 gan2 mat1je5*? *waan2 pe1paai2*; What did the girl have in her hair? *a leaf*; *gam2 kei4sat6 hai6 mat1je5? syu6jip6*) or the "false" component of the sentence (i.e., What did the teacher *think* the girl was doing? *reading*; *gam2 sin1 saang1 ji5wai4 mat1je5? tai2 syu1*; What did the teacher *say* the girl had in her hair? *a bug*; *gam2 sin1 saang1 waa6 mat1je5 bei2 go3 neoi5 zai2 teng1? Jau5 zek3 gwan1cung4 hai2 tau4faat3 dou6*), with equal numbers of test questions for each component. Following de Villiers's procedure, we did not ask the different types of questions for the same sentences, but the pairing of question type to verb and sentence was counterbalanced across children.

All responses were audio-recorded and transcribed, with a score of 1 for each correct response. As with de Villiers and Pyers (2002), we scored the children's responses leniently such that they did not have to report back the entire complement structure. Minimal responses such as those in the examples above were sufficient to demonstrate that they understood and remembered the structure. The total score in this task was 8, and the assignment of sentences to conditions was counterbalanced across children.

It is important to note that de Villiers (de Villiers & de Villiers, 2000; de Villiers & Pyers, 2002) argued that it is the questions about the false components of complements that should be prerequisite to mastery of FB because only these components truly demonstrate that the child is able to linguistically integrate the entire sentence and maintain a representation of the false aspect of the sentence in the face of a differing reality. Nonetheless, given that the relevance of this may differ in a language for which complements do not share the finite syntax that is necessary to talk about beliefs in English, we examined children's responses to both the false and the true (reality/control) components of these sentences.

*FB.* FB understanding was assessed by two tasks that were similar to the change of location task established by Wimmer and Perner (1983) and the unexpected contents task developed by Hogrefe, Wimmer, and Perner (1986). In the change of location task, the children were told a short story in which a character came to hold an FB about the whereabouts of an object. The children then predicted where the character would look for that object. In this story, Yan Yan put the ball in the blue box and then she left. Kai Kai, a friend of Yan Yan, then removed the ball from the blue box to the green one. Yan Yan returned and wanted the ball. Two test questions were asked: "Where does Yan Yan *think* her ball is? Will she *think* the ball is in the blue box or the green one?" (*Gam2, jan1jan1 ji5wai4 keoi5 go3 bo1bo1 hai2 bin1dou6 ne1? Keoi5 wui5 ji5wai4 go3 bo1bo1 hai2 laam4sik1 hap2 leoi5min6, ding6hai6 hai2 luk6sik1 hap2 leoi5min6?*) and "Where will she *look for* the ball? Will she *go to* the blue box or the green one?" (*Gam2 jan1jan1 wui5 heoi3 bin1dou6 wan2 keoi5 go3 bo1bo1 le1? Keoi5 wui5 heoi3 laam4sik1 hap2 go2dou6, ding6 hai6 luk6sik1 hap2 go2dou6?*). The control question "Where *is* the ball now? Is the ball really *in* the blue box or the green one?" (*Gam2 go3 bo1bo1 zan1 hai6 hai2 bin1dou6? hai2 laam4sik1 hap2 leoi5min6*, ding6 hai6 *hai2 luk6sik1 hap2 leoi5min6?*) was asked to make sure children were paying attention to the story.

In the unexpected contents task, children were shown a cylindrical Smarties box (Smarties are commonly eaten candies in Hong Kong) and asked what they thought was inside the box. The experimenter then opened the box to reveal pencils and asked the control question "What does the box really contain?" The experimenter then closed the box and asked, "What did you think the box contained before I opened it? Did you think that it contained candies or pencils?" and (after Big Bird was brought in) "Big Bird has never seen inside the box, so what does Big Bird think is inside the box?"

The maximum score in each task was 2. A score of 1 was given for each correct response to the test questions (i.e., blue box in the change of location task and chocolate in unexpected contents task). The order of questions and choices within the questions were counterbalanced across children. The total FB score was the sum of the unexpected contents task and the change of location task and ranged from 0 to 4.

## *Results*

Descriptive statistics of memory for complements and FB tasks are presented in Table 1. Both memory for complements and FB understanding showed increases in performance with age. FB performance differed across the different age groups, but not across questions or tasks, and all of the children in this sample passed the control questions, regardless of age. Thus, we collapse across question and task and include all children in the subsequent analyses.

The first set of analyses we ran involved a repeated measures analysis of variance (ANOVA) comparing children's performance on the different types of verbs and the different test questions in the memory for complements task, with age group (4- vs. 5- vs. 6-year-olds) and gender as between-subjects variables. Overall, as expected, the older children performed better on the complements task than the younger children, $F(2, 66) = 6.07$, $p < .005$, $pr\eta^2 = .15$, with the 4-year-olds performing significantly more poorly than both the 5-year-olds and the 6-year-olds (Tukey's *p*s $< .05$). There were no main effects or interactions for gender. As can also be seen from the means in Table 1 (collapsed across gender), there was a clear effect of question type, $F(1, 66) = 56.54$, $p < .001$, $pr\eta^2 = .46$, such that the questions about the true components of the sentences were much easier for the children to answer than the false components, as was expected. However, although the difference in the type of verb (mental state vs. communication) used in these sentences did not reach significance, there was a Verb $\times$ Question Type interaction, $F(1, 69) = 11.76$, $p < .001$, $pr\eta^2 = .15$, such that the *think* (*ji5wai4*) sentences were actually easier than the *say* (*waa6*) sentences in the false complement condition,

Table 1
*Study 1 Means and Standard Deviations by Age Group for Age, False Belief, and Complement Tasks*

| | 4-year-olds | | 5-year-olds | | 6-year-olds | |
|---|---|---|---|---|---|---|
| Task | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Mean age | 4.48 | 0.30 | 5.40 | 0.34 | 6.18 | 0.15 |
| Memory for complements | | | | | | |
| *Say* true/control | 1.58 | 0.71 | 1.79 | 0.41 | 1.71 | 0.62 |
| *Think* true/control | 1.37 | 0.65 | 1.58 | 0.50 | 1.75 | 0.44 |
| True Control Total | 2.96 | 1.08 | 3.38 | 0.71 | 3.46 | 0.83 |
| *Say* false complements | 0.42 | 0.65 | 0.79 | 0.83 | 1.00 | 0.83 |
| *Think* false complements | 0.75 | 0.68 | 1.08 | 0.83 | 1.33 | 0.87 |
| False complements total | 1.17 | 1.05 | 1.88 | 1.48 | 2.33 | 1.63 |
| Total Complement score | 4.13 | 1.36 | 5.25 | 1.60 | 5.79 | 2.04 |
| False belief | | | | | | |
| Unexpected contents | 0.50 | 0.72 | 1.00 | 0.72 | 1.25 | 0.74 |
| Change of location | 0.79 | 0.78 | 0.79 | 0.78 | 1.13 | 0.85 |
| Total False Belief score | 1.29 | 1.30 | 1.79 | 1.10 | 2.33 | 1.37 |

*Note.* For all age groups, $n = 24$.

paired $t(71) = 3.94$, $p < .001$, as can be seen in Figure 1. Although this was unexpected given that the English studies never found differences between communication and mental state verbs, it is consistent with Tardif et al.'s (2004) findings that Cantonese-speaking children do better on tests of FB if given the marked "falsely think" verb in the test question. Nonetheless, for both verbs in this study, asking children about the false portion of the sentence was harder than asking them about the true/reality control portion of the sentence, paired $ts(71) = 7.92$ and 4.59, $ps < .001$, for *say* and *think*, respectively.

Our second set of analyses compared children's accuracy levels on the two FB tasks (transfer and unexpected contents), with age and gender as between subjects variables. In this analysis, we found a clear effect of age, $F(2, 66) = 4.45$, $p < .05$, $pr\eta^2 = .12$, but no other main effects or interactions. As expected, the older children performed better than the younger children on these tasks, with a significant difference between the 4- and the 6-year-olds (Tukey's $p < .005$).

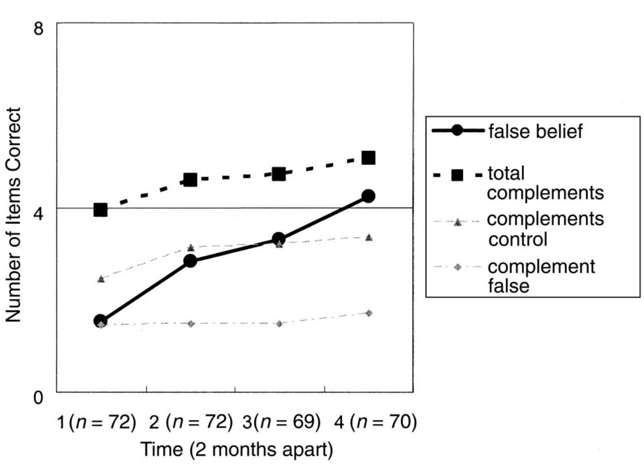



*Figure 1.* Growth trajectory of false belief understanding and memory for complements for Time 1 to Time 4 in Study 2.

Given that we found expected increases in age for both the complements and the FB tasks, we then went on to explore further de Villiers's (1995) hypothesis that complements, and false complements in particular, might be related to children's understanding of FB, even in Cantonese. Specifically, we were interested in finding out whether children who had more advanced complement understanding also scored higher on FB tasks, and thus we conducted a series of correlations to explore this relationship. Overall, and as can be seen in Table 2, there was a significantly positive relationship between performance on the memory for complements task and that on the FB tasks, and this correlation was still significant after age (real age, as calculated from children's birthdates at time of testing) was partialed out. Moreover, consistent with de Villier's hypothesis and with previous findings from studies of English-speaking children, this relationship was more evident for questions about the false components of the complements than it was for true components, and it was consistent for both the *say* and *think* complements. Thus, even for a Cantonese-speaking sample in which one might expect not to find such a relationship, false complement acquisition is associated with FB understanding, even after controlling for age. However, when the three age groups of children were examined separately, the correlation between FB and complements (total and false complements, and both *say* and *think* false complements) remained only for the 5-year-olds. This is interesting because, in our sample, this is the period at which children went from below-chance performance to beginning to show some evidence of FB understanding. These data also suggest that a longitudinal study targeting children who start out at below-chance performance on FB and following both their complement and FB understanding through to a point at which it reaches or exceeds chance

Table 2
*Correlations Between False Belief and Complement Measures in Study 1*

| Variable | False belief | False complements | Total complements |
|---|---|---|---|
| Whole sample ($n$ = 72) simple/partial (age) correlations | | | |
| True complements | .21/.11 | .11/.01 | .59**/.35** |
| *Say* false complements | .40**/.34** | .90**/.90** | .77**/.53** |
| *Think* false complements | .36**/.28* | .91**/.90** | .81**/.75** |
| False complements | .42**/.34** | — | — |
| Total complements | .45**/.35** | .87**/.85** | — |
| 4-year-olds ($n$ = 24) | | | |
| True complements | .07 | −.19 | .65** |
| *Say* false complements | .21 | .78** | .33 |
| *Think* false complements | .14 | .80** | .65** |
| False complements | .22 | — | — |
| Total complements | .22 | .62** | — |
| 5-year-olds (n = 24) | | | |
| True complements | .16 | −.08 | .37 |
| *Say* false complements | .57** | .89** | .79** |
| *Think* false complements | .54** | .89** | .81** |
| False complements | .62** | — | — |
| Total complements | .65** | .90** | — |
| 6-year-olds ($n$ = 24) | | | |
| True complements | .24 | .30 | .65** |
| *Say* false complements | .27 | .96** | .92** |
| *Think* false complements | .23 | .96** | .85** |
| False complements | .26 | — | — |
| Total complements | .31 | .92** | — |

* $p < .05$. ** $p < .01$.

performance would be a particularly good strategy for examining this relationship.

To conclude, the results in Study 1 showed that children who had better complement understanding performed better on FB tasks, even in Cantonese. Moreover, even though these Cantonese-speaking children performed better on the *think* false complements than on the *say* false complements, performance on both types of sentences was equally related to FB understanding. Although finding a relationship between these two concepts at all in Cantonese is somewhat surprising, it still does not tell us anything about whether complement understanding is indeed a precursor to FB, as suggested by de Villiers (1995), nor does it tell us whether there may be underlying mechanisms that mediate or explain this relationship.

Thus, in our second study, we preselected Cantonese-speaking children who were at the cusp of and had not yet achieved FB understanding and followed them longitudinally to examine their acquisition of both complement and FB understanding. Moreover, we designed our study such that we could use hierarchical linear modeling (HLM) techniques to help us examine the developmental trajectory for FB in Cantonese. Although traditional regression techniques can help in the untangling of causal relations across time when one puts two highly related variables in parallel models, with one as a predictor and the other as an outcome across time, these techniques are cumbersome when multiple variables and multiple time points are involved. In addition, one can only examine pairs of time points (Time 1 to Time 2, Time 1 to Time 3, etc.) under this type of modeling; and when a variable has an effect in one pairing but not in another, it is difficult to interpret the impact of that variable in a larger sense. Finally, a major advantage to HLM and other growth modeling techniques is that they allow us to plot growth curves and to understand how variables might affect either an intercept value, the rate of growth (slope), or both. Multiple regression techniques are useful for modeling the effects of a set of covariates on an outcome, but they are not helpful for modeling actual growth trajectories of that outcome. However, because previous longitudinal studies used only multiple regression techniques, we present our data in both formats and try to show how the use of HLM is both more parsimonious and more meaningful for answering causal questions about the development of multiple variables over time.

## Study 2

In addition to relationships between complements, or language more generally, and FB understanding, other researchers have argued that specific cognitive factors, namely executive functioning abilities, are predictive of FB understanding. In particular, working memory and inhibitory control (Carlson et al., 2002) are two aspects of executive functioning that have shown clear relations to FB performance, even when language abilities have been controlled (Carlson, Mandell, & Williams, 2004; Hughes, 1998b). As with language skills, autistic children perform poorly on tests of executive functioning, and their ability to perform well on tests of executive functioning is directly related to their ability to perform FB tasks (Hughes & Russell, 1993). Moreover, training on inhibitory control skills such as those required to do well in Luria's (1965) hand game has been shown to have a direct impact on FB performance in a normally developing sample (Flynn, O'Malley, & Wood, 2004; Kloo & Perner, 2003).

Although executive functioning can be and is conceived of in a number of ways, there are two components that we believe are particularly relevant for understanding children's developing abilities to understand both FBs and complements—working memory and inhibitory control. Working memory is particularly relevant in that children with larger working memories will be able to hold more, and perhaps contradictory, information in their minds. Given that both FB tasks and sentence complements require one to hold two, generally contradictory, pieces of information in mind in order to resolve and respond to a test question, it is logical that individual differences in working memory might account for individual differences in both of these tasks. At the same time, being able to simply hold these pieces of information in mind is not enough. Children could still report the true state of affairs, even though they know both states, when answering FB prediction questions simply because they are unable to inhibit a "reality bias" in responding to a tester's questions (Cassidy, 1998; Mitchell, 1994). Therefore, some ability to inhibit a prepotent response is also needed.

Indeed, in a study examining the relationship between executive functioning and FB understanding, Hughes (1998a, 1998b) found a significant effect of both of these factors, and a combined effect remained even when age, initial FB performance, and language abilities were controlled for. Not only was executive functioning correlated with FB understanding but it also predicted performance on theory of mind tasks 1 year later (Hughes, 1998b). It is interesting to note that the relationship was not symmetrical. That is, early executive functioning predicted theory of mind performance but not vice versa. In Kloo and Perner's (2003) study, however, training on either executive function or FB influenced the development of the other.

Thus, our goals for Study 2 were twofold. First, we wanted to examine the relationship between FB and complement understanding more carefully in a longitudinal sample of Cantonese-speaking children to examine the critical "complement as precursor" aspect to de Villiers's (1995) hypothesis that we could not examine in our cross-sectional sample. Second, we wanted to determine whether executive functioning skills, together with more general linguistic abilities, can help explain the relationship between FB and sentence complements in Cantonese. It is possible that although general language abilities themselves may not fully account for the relationship (as de Villiers & Pyers, 2002, and Hale & Tager-Flusberg, 2003, found), the combination of language abilities with executive functioning abilities may explain the relationship, and finding evidence for this in a Cantonese-speaking sample would certainly provide strong justification for reexamining the relationship in English.

## Method

### Participants

Children in Study 2 were selected from the same kindergartens and had the same general demographic profiles as the children in Study 1. At Time 1, 81 Cantonese-speaking preschool-aged children who did not participate in Study 1 were tested and screened for FB and complement understanding. Among this initial group of children were 12 three-year-olds ($M = 3.58$, $SD = 0.16$), 53 four-year-olds ($M = 4.45$, $SD = 0.37$), and 16 five-year-olds ($M = 5.38$, $SD = 0.25$). Given that our design was targeted at finding children who could not yet understand FB and at tracking the development of this understanding and whether complement understanding would precede FB understanding, we then chose the 72 children who "failed" the FB tests (but who "passed" the FB control questions) by scoring two marks or less out of four at this initial screening. Of these, 12 children were 3 years old ($M = 3.58$, $SD = 0.16$), 48 children were 4 years old ($M = 4.33$, $SD = 0.26$), and 12 children were 5 years old ($M = 5.5$, $SD = 0.19$), with ages again calculated from children's actual birthdates.

The 72 selected children were then followed for three additional testing points spaced approximately 2 months apart. Of these 72 children, 3 children failed to show up (1 was absent from school and 2 transferred to other schools) at either Time 3 or Time 4, and an additional 8 children scored full marks on either the complements or the FB tasks at Time 2 or Time 3. The 3 children who did not show up for one or more testing points were dropped from the longitudinal analyses. In contrast, in order to conserve our testing efforts and avoid repeated testing of children who consistently gave correct answers, the 8 children who scored full marks (4 on FB, 1 on complements, and 3 on both complements and FB) were assigned full scores on both complements and FB tasks at Time 3 and/or Time 4.[2]

### Design, Materials, and Procedure

At Time 1, only FB and the memory for complements tasks were administered. These tasks were identical to those we used in Study 1.

Starting from Time 2, we administered four FB tasks (two unexpected contents and two change of location tasks) and one memory for complements task. The scenarios of these tasks were designed to be parallel in structure but different in content at each time point. For the change of locations, for instance, instead of the characters KeiKei and YanYan and a ball with two boxes, the children were told stories about other characters; the moved objects were a book, a chocolate bar, or a bag of potato chips; and locations were brother and sister's rooms, cupboards versus drawers, or refrigerators versus cupboards. Similarly, for the unexpected contents tasks, instead of a Smarties box and pencils, the

[2] Note that we did an additional set of both the regression and the HLM analyses by using only the 59 children for whom we had full data, and this did not change which variables became significant in any of our analyses.

objects included a crayon box and bandaids, a cookie container and candies, and a potato chip box and a ruler. The composite score for FB understanding at each time point, was 8 (2 test questions × 2 tasks × 2 different types of tasks). Cronbach's alphas for the multiple FB questions at each time point were 0.84, 0.99, and 0.99 for Times 2–4, respectively. Reliability at Time 1 was not meaningful because we selected children who were performing at floor on this measure.

The memory for complements tasks were changed in similar ways as were the FB tasks—the content was changed, whereas the structure was preserved. In one sentence, for example, the teacher said/thought there was a bug in the girl's hair, but it was just a leaf, whereas in the parallel version, a friend said/thought there was spaghetti on the plate, but really it was string. As with Study 1, there were eight sentences presented at each time point, with half of the questions asking about the contents of the false portion of the sentence (i.e., what was said/thought) and half of the questions asking about the real state of affairs. Given that there were no differences in how well the *say* and *think* complements predicted FB understanding in Study 1, we collapsed across these two verbs to create a *memory for complements* score by using the questions about the false components of the complements with a maximum score of 4 for each time point. Cronbach's alphas for the false complements were 0.51, 0.76, 0.79, and 0.77 for Times 1–4. Note that although the reliability of this task was somewhat lower than that of the FB measures, it was computed on fewer items (four questions instead of eight), across two types of verbs (*say* vs. *think*), and is still well within acceptable limits.

At Time 2, we also measured children's short-term memory, inhibition, general language abilities, and nonverbal intelligence.

*Short-term memory.* There were two short-term memory tasks, a digit span task and a noisy book task, both of which showed positive correlations with FB in Hughes's (1998a, 1998b) examination of the relationship between executive function and ToM abilities for this age group. The order of these two tasks was counterbalanced across children, with no significant order effects in children's performance on the tasks.

For the digit span task, the experimenter said a series of single-digit numbers, presented one per second, and the child was asked to immediately recall the digits spoken by the experimenter in the order they were presented. Children were first given two practice trials with two single-digit numbers before testing began. If children failed one or both of the practice trials, the experimenter repeated the instructions and presented another set of practice trials. None of the children in our sample failed this second set of practice trials. Testing began with three trials of two single-digit numbers. If children could recall all the digits presented in at least two of the three trials, they were presented another three trials for three digits, then four, and so on to a maximum of eight digits. Testing was terminated when the child failed two or more trials at any level.

The noisy book task (Goldman, Fristoe, & Woodcock, 1974) was similar to the digit span task except that children were required to recall the animal labels presented in a "noisy book." In our study, this was a toy car which had eight different buttons that had the picture and sound effect of a different animal. Children were initially asked to name each picture as he or she pressed the button to find out what noise it made. This allowed the experimenter to refer to the pictures in the same way as the child did (e.g., if the child called a panda a "bear," then the experimenter did too). If the child failed to name the picture, then the experimenter labeled the picture and asked the child if he or she could press the picture corresponding to the label given by the experimenter. Once the experimenter was certain the child knew the names of the animals, she then went on to present the practice trials. For the practice trials, the experimenter told the child that she would cover up the buttons with a piece of paper and say the names of two of the animals. Then she removed the paper and asked the child to find the pictures for the animals she labeled. The practice trial procedure was the same as on the digit span task, and testing began with three trials of two animal labels, followed by three animals, and so on to a maximum of eight animals. Testing was discontinued if children failed to recall two or more labels in each set of three trials. The scoring procedures were the same as those in the digit span task. The total score was determined by the maximum number of labels children could recall. The highest score was 7.

To improve reliability and reduce the overall number of variables in our regression and HLM analyses, we converted raw scores on each of the subtests to standardized $z$ scores and combined the $z$ scores to form an aggregate measure of short-term memory abilities. As can be seen from Tables 3 and 4, although the means of these two tasks were quite different (albeit very similar to the means for children in Hughes's [1998a, 1998b] study for each of the tasks), the correlations between tasks were moderately strong, thus justifying an aggregate measure.

*Inhibition.* Children's abilities to inhibit prepotent responses were assessed by two independent tests commonly used as measures of inhibitory control in executive functioning batteries (e.g., Diamond & Taylor, 1996; Hughes, 1998a, 1998b). The two tasks chosen for this study were Luria's hand game (Luria, 1965) and a Stroop-like task (Gerstadt, Hong, & Diamond, 1994), both of which have been shown to have moderately positive relations to FB performance in previous studies (Carlson et al., 2004; Hughes, 1998b). In our study, as in most other studies that have used these tasks, both tasks had a control/imitative condition, presented first, and a conflict/test condition, presented after the control condition for all children. As with the two short-term memory tasks, the order of the two inhibition tasks (Luria's, Stroop) was counterbalanced across children, and there were no significant order effects.

For Luria's hand game, the experimenter asked the child to imitate one of two possible gestures (pointing one finger or pointing two fingers) that were intermingled and randomly arranged into a series of 15 trials that were the same for all children in the study. In the control condition, if the experimenter pointed one finger, the child was asked to point one finger. In the conflict condition, the child was asked to make the opposite gesture as that made by the experimenter. For example, if the experimenter pointed with one finger, the child should point with two fingers and vice versa. Before testing began, two practice trials were given in each condition. Children received feedback during the practice trials and were corrected if they made mistakes, and then testing began. In each condition, testing ended either when the maximum number of trials had been reached (15) or when the child performed

Table 3
*Descriptive Statistics for All Variables in Study 2*

| Variable | Time 1 ($n$ = 72) | | Time 2 ($n$ = 72) | | Time 3 ($n$ = 70) | | Time 4 ($n$ = 71) | |
|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| Age | 4.39 | 0.61 | 4.65 | 0.61 | 4.90 | 0.61 | 5.15 | 0.61 |
| FB[a] | 1.52 | 1.62 | 2.83 | 2.65 | 3.31 | 2.79 | 4.23 | 2.68 |
| Complements (Total)[b] | 3.93 | 1.57 | 4.60 | 1.95 | 4.71 | 1.82 | 5.06 | 1.96 |
| False complements | 1.46 | 1.24 | 1.47 | 1.36 | 1.49 | 1.54 | 1.71 | 1.57 |
| True (control) question for complements | 2.47 | 1.28 | 3.12 | 1.28 | 3.23 | 1.00 | 3.35 | 1.14 |
| Short-term memory composite | | | 8.18 | 2.47 | | | | |
| STM-digit span | | | 5.26 | 1.67 | | | | |
| STM-noisy book | | | 2.92 | 1.47 | | | | |
| Nonverbal IQ | | | 11.61 | 3.28 | | | | |
| Inhibition composite (Luria recoded) | | | 15.39 | 5.05 | | | | |
| Inhibition day/night Stroop control (*N* correct) | | | 8.83 | 2.58 | | | | |
| Inhibition day/night Stroop test (*N* correct) | | | 7.20 | 3.57 | | | | |
| Inhibition Luria control (trials to criterion) | | | 6.31 | 1.38 | | | | |
| Inhibition Luria test (trials to criterion) | | | 7.49 | 3.22 | | | | |
| Reynell–Total | | | 110.43 | 13.25 | | | | |

[a] False Belief (FB) Scores at Time 1 were actually based on a total of four test questions, whereas they were based on eight test questions at Times 2, 3, and 4. In order to equate across time periods, the scores were doubled in Time 1 for comparison purposes only. [b] In the reported analyses, data were imputed for 6 participants at Time 3 and for 8 participants at Time 4, who had a perfect score on FB at Time 2 or Time 3 in hierarchical linear modeling analyses. For these participants, perfect scores assumed for complement and FB tasks in descriptive means, correlations, and regression analyses. However, it is important to note that we also ran the entire set of analyses with just the subset of children on whom we have full data and neither the overall pattern of results, nor the significant findings, changed. We report the fuller set of data for completeness.

correctly on 6 consecutive trials. Scoring for each condition was based on the number of trials to criterion (i.e., the number of trials required to score 6 in a row correct, with a maximum of 15 for each condition and lower scores reflecting higher inhibition abilities). These scores were then converted and reversed by subtracting the number of trials to criterion from 16 in order to create a maximum of 10, with higher scores indicating better performance, in order to create a combined inhibition score with the Stroop-like task.

The day–night Stroop task was similar to Luria's hand game in that it had both a control and a conflict condition. At the beginning, children were shown the "sun" card and the "moon" card and asked to say when each of these appeared. In the control condition, the experimenter asked the children to say "day" when shown the sun card and to say "night" when shown the moon card. In the conflict condition, children were asked to say the opposite—"night" for sun and "day" for moon. In this task, there were 2 practice trials and 10 test trials in each condition. Feedback was given and instructions were repeated if children failed the practice trials. The presentation order of sun and moon cards was identical across children. Scoring in this task was based on the number of trials correct in each condition, with a maximum of 10 per condition.

Raw scores from both the control and conflict (test) components of these two measures of inhibitory were converted to $z$ scores and then combined to form an aggregate measure of inhibitory skills. Again, the correlations between tasks and components of tasks, as shown in Table 4, were moderately strong and justified the aggregation so that we could reduce the total number of variables entered into our analyses.[3]

*General language abilities.* This was assessed by the Reynell Developmental Language Scale (Reynell & Huntley, 1985; Hong Kong Version, 1987), which is the only comprehensive measure of language abilities normed for this age group in Hong Kong. The scale is normed for ages 1.0 to 7.0 and includes several subscales, organized into two main parts. For the Receptive (comprehension) part, seven subscales measuring progressively more difficult understandings of vocabulary and syntax are administered, with a total of 67 items. For the Expressive portion, subscales exist for language structure (23 items including a range of structural questions from phonology to pragmatics), vocabulary (26 items, but with vocabulary target items embedded in complex syntactic frames), and ability to verbalize connected thoughts creatively (24 items). A speech therapist trained specifically in the administration of the Reynell conducted and scored the audio-recorded transcripts from this task. Note that for the purposes of our study, only the total (raw, not age-normed) scores (maximum = 140) were used because none of the subscales cleanly separates syntax from vocabulary or other components of language.

*Nonverbal intelligence.* The pattern analysis subtest of the Stanford–Binet Intelligence Scale (Thorndike, Hagan, & Sattler, 1986) was used to assess nonverbal ability. This scale is normed in Hong Kong for this age range and shows excellent psychometric

[3] It is important to note that we did not combine the short-term memory and inhibition measures. Our primary reason for this is that they were only moderately correlated, and we were especially interested in examining whether they showed independent contributions to predicting FB performance or to moderating the predictability of false complements on FB. However, we also examined models in which the inhibition and short-term memory measures were combined to form an aggregrate executive function score and found that the aggregated measure resulted in diminished, not heightened, correlations with both our FB and false complements measures.

Table 4
*Simple Correlations (Top Half of Diagonal), and Partial Correlations (Age, Bottom Half of Diagonal) for All Variables in Study 2*

| Variable | FB1 | FB2 | FB3 | FB4 | CTOT1 | CTOT2 | CTOT3 | CTOT4 | FC1 | FC2 | FC3 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Age1 | −.05 | .33** | .36** | .28* | .23 | .45** | .45** | .44** | .11 | .30* | .38** |
| FB1 | — | .29* | .33** | .36** | .38** | −.04 | .22 | .29* | .24* | .00 | .14 |
| FB2 | .35** | — | .74** | .61** | .30* | .39** | .55** | .60** | .23 | .28* | .58** |
| FB3 | .37** | .70** | — | .80** | .37** | .35** | .60** | .65** | .25* | .30* | .61** |
| FB4 | .39** | .57** | .81** | — | .31** | .37** | .48** | .61** | .14 | .28* | .46** |
| Ctot1 | .41** | .22 | .31* | .25* | — | .18 | .44** | .45** | .61** | .04 | .31** |
| Ctot2 | −.03 | .25* | .21 | .27* | .06 | — | .43** | .36** | .11 | .76** | .31** |
| Ctot3 | .27* | .48** | .53** | .42** | .39** | .29* | — | .73** | .32** | .39** | .84** |
| Ctot4 | .34** | .54** | .59** | .56** | .41** | .20 | .67** | — | .26* | .32** | .66** |
| FC1 | .24* | .20 | .22 | .12 | .61** | .03 | .30* | .24 | — | −.05 | .14 |
| FC2 | .01 | .20 | .21 | .22 | −.06 | .71** | .30* | .22 | −.09 | — | .43** |
| FC3 | .17 | .52** | .55** | .40** | .25* | .17 | .81** | .60** | .11 | .30** | — |
| FC4 | .32** | .47** | .50** | .46** | .40** | .10 | .55** | .78** | .23 | .28* | .62** |
| RTOT | .04 | .29* | .51** | .51** | .24* | .20 | .45** | .52** | .04 | .29* | .51** |
| IQ | .06 | .10 | .18 | .03 | .26* | .04 | .04 | .17 | .06 | .10 | .18 |
| STM-DIG | .15 | .21 | .17 | .32** | .17 | .48** | .21 | .29* | .05 | .34** | .08 |
| STM-BK | −.04 | .11 | .15 | .26* | .06 | .19 | .21 | .18 | .07 | .06 | .15 |
| STM-TOT | .07 | .21 | .20 | .37** | .15 | .44** | .27* | .30* | .08 | .26* | .15 |
| ST-TEST | −.04 | .04 | .22 | .06 | .16 | .15 | .28* | .26* | .16 | .17 | .15 |
| ST-TOT | −.05 | −.02 | .23 | .16 | .15 | .06 | .29* | .27* | .19 | .02 | .14 |
| LU-TEST | .01 | .20 | .11 | .12 | .08 | .23 | .08 | .17 | .07 | .01 | .03 |
| LU-TOT | .01 | .18 | .13 | .17 | .15 | .21 | .07 | .21 | .11 | .01 | .01 |
| INH-TEST | −.02 | .17 | .23 | .12 | .16 | .26* | .25* | .30* | .16 | .12 | .12 |
| INH-TOT | −.02 | .10 | .22 | .21 | .18 | .17 | .22 | .30* | .18 | .02 | .09 |

*Note.* When 1, 2, 3, or 4 follows an abbreviation, it indicates the time of testing (i.e., Time 1, 2, 3, or 4). FB = false belief; C = complement; TOT = total; FC = false complement; R = Reynell; STM = short-term memory; DIG = digits task; BK = noisy book task; ST = Stroop; LU = Luria; INH = inhibition.
* $p < .05$. ** $p < .01$.

properties, with alphas over .90 in the age range, high correlations with other aspects of nonverbal IQ, and moderate correlations with overall IQ. Thus, we considered it to be an appropriate measure of nonverbal reasoning skills/intelligence for our sample. The first part of this test involved six trials in which the experimenter presented children with a form board with the pieces in place, removed the pieces in front of them, and asked children to put the pieces back into the holes. The first items did not change the orientation of the pieces, whereas later items got progressively harder by either adding pieces and/or changing their orientation so that they no longer matched the model. In the second part, the child was presented with a number of identical cubes, each with six different faces. They were then asked to arrange the cubes such that the top faces of the cubes were identical to the structure produced by the experimenter. These items also increased in difficulty, with greater and greater numbers of cubes and more complex designs in successive trials. There were a total of 12 trials and a time limit of 30 s/trial. The highest mark children could get on this test was 18.

## *Results*

Mean levels and descriptive statistics for variables at all time points are presented in Table 3. As expected, the FB composite score, which was the sum of scores in the unexpected contents and change of location tasks, increased steadily across the four time points, and such change over time appeared to be linear, as shown in Figure 1.

Moreover, both the simple and age-partialled (real age at Time 1, as calculated from children's birthdates) correlations at each testing point (i.e, Time 1 FB with Time 1 complements) show a strong relationship between children's performance on the FB and memory for false complements task, thus replicating our findings from Study 1, as shown in Table 4. As is also evident from Table 4, both FB understanding and memory for complements were highly associated with age, replicating several findings in English and our own findings in Study 1 (Astington & Jenkins, 1999; Happé, 1995; Jenkins & Astington, 1996). In addition, general language abilities were highly associated with FB understanding at all time points, which again replicates findings reported repeatedly in the literature (Astington & Jenkins, 1999; Cheung et al. 2004; de Villiers & Pyers, 2002; Peterson & Siegal, 1995). Somewhat surprisingly, however, the relations between short-term memory, inhibition, and nonverbal IQ were inconsistently correlated with FB understanding, and these correlations tended to disappear when controlling for the child's age. A similar pattern was found with the memory for complements task, with general language abilities showing more consistent correlations than any of the other control variables.

Overall, then, simple correlations between memory for complements and FB understanding exist for both cross-sectional and longitudinal samples of Cantonese-speaking children. However, correlation results alone are not sufficient to provide evidence for de Villier's (1995) strong hypothesis that children's ability to master complements, and particularly false complements, predicts

| FC4 | RTOT | IQ | STM-DIG | STM-BK | STM-TOT | ST-TEST | ST-TOT | LU-TEST | LU-TOT | INH-TEST | INH-TOT |
|---|---|---|---|---|---|---|---|---|---|---|---|
| .36** | .59** | .54** | .24* | .36** | .38** | .20 | .34** | .24* | .26* | .32** | .38** |
| .28* | −.03 | .03 | .11 | −.05 | .04 | −.06 | −.16 | .02 | .04 | −.03 | −.07 |
| .53** | .44** | .23 | .30* | .22 | .33** | .12 | .01 | .29* | .26* | .28* | .17 |
| .57** | .59** | .33** | .25* | .26* | .32** | .27* | .18 | .22 | .24* | .33** | .26* |
| .51** | .55** | .17 | .36** | .33** | .43** | .10 | .13 | .20 | .27* | .21 | .25* |
| .44** | .31** | .32** | .18 | .16 | .22 | .17 | .19 | .12 | .21 | .22 | .25* |
| .25* | .44** | .24* | .52* | .28* | .51** | .27* | .23* | .27* | .18 | .38** | .26* |
| .62** | .59** | .27* | .28* | .34** | .40** | .33** | .31* | .19 | .21 | .36** | .32** |
| .81** | .62** | .37** | .34** | .31* | .41** | .30* | .24* | .27* | .35* | .40** | .36** |
| .26* | .59** | .54** | .10 | .11 | .13 | .20 | .25* | .07 | .11 | .19 | .22 |
| .36** | −.03 | .03 | .41** | .13 | .34** | .26* | .14 | .07 | −.00 | .22 | .09 |
| .67** | .44** | .23 | .15 | .28* | .27* | .20 | .16 | .13 | .15 | .23 | .19 |
| — | .59** | .33** | .18 | .08 | .16 | .32** | .21 | .14 | .23 | .31** | .28* |
| .51** | — | .47** | .50** | .53** | .66** | .37** | .51** | .38** | .43** | .53** | .59** |
| .03 | .29* | — | .19 | .34** | .34** | .19 | .20 | .29* | .34** | .34** | .34** |
| .12 | .44** | .13 | — | .23* | .79** | .31** | .24* | .26* | .27* | .39** | .32** |
| −.06 | .42** | .19 | .18 | — | .79** | .10 | .25* | .29* | .37** | .28* | .39** |
| .04 | .56** | .20 | .77** | .77** | — | .26* | .32** | .35** | .41** | .42** | .45** |
| .29* | .32* | .15 | .24 | .05 | .19 | — | .87** | .08 | .12 | .73** | .58** |
| .25* | .42** | .08 | .19 | .17 | .24 | .87** | — | .20 | .27* | .77** | .80** |
| .04 | .33** | .19 | .23 | .23 | .29* | .04 | .20 | — | .92** | .74** | .69** |
| .11 | .40** | .20 | .29* | .28* | .37** | .12 | .31* | .93** | — | .73** | .80** |
| .23 | .45** | .24 | .32** | .19 | .34** | .72** | .75** | .72** | .74** | — | .91** |
| .22 | .50** | .18 | .30* | .28* | .38** | .60** | .79** | .72** | .82** | .91** | — |

and precedes their ability to understand FB. In order to evaluate this hypothesis more directly, we need to consider the developmental course of both complement and FB understanding in our developmental sample. Moreover, we need to consider the role of other factors, such as general language abilities or short-term memory, which may mediate or completely account for this relationship. In order to explore these issues, we use two analysis strategies to combine our multiple independent variables within our longitudinal design—a hierarchical series of regression analyses and HLM. We chose the first analysis method to provide a straightforward comparison with previous research on English-speaking children for our Cantonese-speaking sample. However, we believe that the second analysis method is superior in modeling and understanding the growth of FB over time and how it may be related to complement understanding and other types of measures such as general language or executive functioning abilities. Nonetheless, because HLM has not been used in the ToM literature and is still underutilized in developmental research, we present both methods. We then compare these results with the aim of achieving a fuller understanding of how complements and FB might be related and how different designs and analysis methods can lead one to different conclusions about the nature of this relationship.

### *Multiple Regression Analyses*

In our multiple regression analyses, we considered a series of models to examine whether or not children's memory for false complements could uniquely predict their performance on FB tasks in our longitudinal design. However, to conserve space, we do not present each of these models in detail because each model requires six parallel analyses (Time 1–Time 2, Time 1–Time 3, Time 1–Time 4, Time 2–Time 3, Time 2–Time 4, Time 3–Time 4). Instead, we present just the first and the final of these models. In addition, although we ran models for both the total complements and false complements scores, we present only the false complements results because de Villiers's (1995) hypothesis specifically focuses on children's ability to understand the false component of a complement. It is important to note, however, that even though the total complements showed clearer evidence of linear growth over our testing period, the overall pattern of results did not change when we used total complements instead of false complements as our predictor variable in either the regression or the HLM analyses.

### *False Complements Predicting FB Over Time, Controlling for Age and Prior FB Performance*

As we saw in Table 4, children's scores on both the false complement component and the complement task overall were related to their FB performance, but this was not always a consistent relationship, nor did it always appear to be predictive such that complement scores at a previous testing session were related to FB scores at a later testing session (only one out of six possible pairings showed a significant correlation). In fact, the opposite relationship, FB scores at previous testing sessions correlated with complement scores at later testing sessions, was true even more often (four out of six pairings showed significant correlations for this pattern). Thus, our first series of models involved adding children's FB performance at the prior testing session to the model as well to see if complements continued to show any predictability in this sample.

As can be seen in Table 5, once children's prior FB performance was added to the model to control for the contemporaneous correlations between complements and FB, false complements failed to predict unique variance in FB performance for any of the time intervals measured. Thus, for Cantonese-speaking children, there is no evidence to support de Villiers's (1995; de Villiers & de Villiers, 2000) hypothesis that complements are prerequisite to understanding FB.

### *False Complements Predicting FB Over Time*

In the next step, we added general language abilities as a control, and, finally, we added the other control variables to see which of the various independent variables (control or false complements) might be significant predictors of FB performance over this time period. For both short-term memory and inhibition,

Table 5
*Hierarchical Regression Analyses Predicting False Beliefs (FB) From False Complements, Controlling for Age and Initial FB Performance in Study 2*

| Step and variable | Unstandardized β | *SE* β | Standardized β | $R^2/\Delta R^2$ |
|---|---|---|---|---|
| Time 1 to Time 2 ($n = 72$) | | | | |
| Step 1 | | | | .20*** |
| Age | 1.50 | .47 | .35** | |
| FB, Time 1 | 0.99 | .35 | .30** | |
| Step 2 | | | | .02 |
| False complements, Time 1 | 0.28 | .24 | .13 | |
| Time 1 to Time 3 ($n = 70$) | | | | |
| Step 1 | | | | .25*** |
| Age | 1.72 | .48 | .38** | |
| FB, Time 1 | 1.20 | .36 | .35** | |
| Step 2 | | | | .02 |
| False complements, Time 1 | 0.30 | .24 | .13 | |
| Time 2 to Time 3 ($n = 70$) | | | | |
| Step 1 | | | | .57*** |
| Age | 0.54 | .39 | .12 | |
| FB, Time 2 | 0.73 | .09 | .70*** | |
| Step 2 | | | | .00 |
| False complements, Time 2 | 0.01 | .16 | .01 | |
| Time 1 to Time 4 ($n = 71$) | | | | |
| Step 1 | | | | .22*** |
| Age | 1.31 | .47 | .30** | |
| FB, Time 1 | 1.24 | .35 | .38** | |
| Step 2 | | | | .00 |
| False complements, Time 1 | 0.05 | .24 | .02 | |
| Time 2 to Time 4 ($n = 71$) | | | | |
| Step 1 | | | | .38*** |
| Age | 0.38 | .44 | .09 | |
| FB, Time 2 | 0.59 | .10 | .59*** | |
| Step 2 | | | | .01 |
| False complements, Time 2 | 0.22 | .18 | .12 | |
| Time 3 to Time 4 ($n = 70$) | | | | |
| Step 1 | | | | .69*** |
| Age | −0.09 | .32 | −.02 | |
| FB, Time 3 | 0.81 | .07 | .84*** | |
| Step 2 | | | | .00 |
| False complements, Time 3 | −0.13 | .16 | −.07 | |

** $p < .01$. *** $p < .001$.

aggregate measures, derived from *z*-scores on individual tasks, were used. For inhibition, both control and conflict trials were included in our total inhibition score, although the actual patterns of predictability from the test versus total or individual versus aggregate scores did not differ.

Despite the fact that memory for complements did not contribute unique variance to FB performance in this sample, individual differences in children's language abilities, as measured by the Reynell, was a significant predictor of FB performance at almost all of the time points measured in this study. It is interesting to note that for the models (Time 3 to Time 4) in which language was not as strong a predictor, children's short-term memory abilities emerged as a predictor, as shown in Table 6. Given this finding and the appearance of short-term memory as a significant predictor in Step 1 for the Time 1 to Time 4 and Time 2 to Time 4 intervals, we also ran a set of analyses which entered Reynell scores at Step 1 and examined whether or not short-term memory would emerge as a unique predictor of FB performance at Step 2. However, this model failed to find a unique role for short-term memory in predicting FB. Finally, we ran a further set of models in which we did not include FB performance at Time 1 as a control variable. In these models, the Reynell again emerged as the most significant predictor of FB performance for each of the time intervals tested, with short-term memory emerging as a significant (for Times 1, 2, and 3 to Time 4) but not a unique predictor of FB. However, it is interesting that when FB at Time 1 was not used as a control variable, false complements emerged as a significant and unique predictor of FB performance in the Time 2 to Time 3 and the Time 3 to Time 4 models. Nonetheless, when we find such patterns at one set of pairings across time and not another, it is difficult to make general statements on the role of these variables in predicting FB performance. For this reason and for the others mentioned above, we feel that a more comprehensive way of examining the effects of complements or any of these other variables on FB performance is to consider the entire trajectory of FB under a single model, and this is exactly what HLM allows us to do.

## *HLM*

In HLM, the general process of hypothesis testing is similar to that used in regression analyses. Sample size requirements for HLM are also relatively similar to those for other multivariate techniques. Specifically, Tabachnick and Fidell (1996) recommended a ratio of 5 to 10 subjects per estimated parameter (our models have between three and eight parameters). The advantages to HLM are that multiple waves of data can be considered simultaneously, and the actual growth functions can be determined and examined within a single model. Moreover, HLM is especially good for describing trajectories of change within individuals over time because of the way in which occasions of measurement (time of testing, Level 1) are nested within individuals (Level 2; Bryk & Raudenbusch, 1987; Schulenberg & Maggs, 2001). Thus, in HLM, unlike the repeated linear regression models considered above, we can carefully examine the growth trajectories of both FB and false complements, first, to determine whether their growth over time is linear and, second, to examine the growth in each of these factors as a function of the other and as a function of the additional language, executive function, and nonverbal intelligence measures we collected.

In the following models, we considered the children's FB and false complement performance at Time 1 to be their "baseline" levels of performance and modeled changes in performance from Time 2 to Time 4.

In order to examine more fully de Villiers's (1995) hypothesis that FB understanding is not just related to, but is also preceded by, false complement understanding, we used children's FB understanding at Times 2, 3, and 4 as our outcome variable and examined a number of hypotheses and models regarding the growth in FB over time. In our models, we considered two levels of analysis. At Level 1, we model overall growth on FB or false complements as a function of time and examine whether this is best expressed as a linear or a quadratic function. At Level 2, we considered individual differences that were not considered to be time-varying—such as initial levels of complement or FB understanding, relative age with respect to other children in the study, language abilities, executive functioning abilities, and nonverbal IQ—to determine whether these individual differences contribute to explaining differences in the trajectories of FB or false complements. Following the tradition of HLM, the slope and intercept parameters from the Level 1 FB scores then become outcome variables at Level 2 so that the effects of these individual difference variables on both the slope and the intercept of FB could be examined directly for the sample as a whole (Bryk & Raudenbusch, 1992).

*Hypothesis 1A:* Linear growth of FB over time.

It is customary in HLM to begin with testing the unconditional (basic growth) model. Thus, we begin with this model and explicitly test whether the growth could be accounted for by linear or quadratic growth curves and to examine whether significant variation in trajectories exist such that it would make sense to continue to examine whether any of the additional variables could account for this variation. By far the best fitting growth model for our data was linear, with no additional variance being explained by the quadratic model and nonsignificant contributions of the quadratic parameter, $t(208) = 0.627$, $p = .530$. Thus, all further models for FB understanding were built from the following linear growth model with time centered such that time = 0 corresponds to our Time 2 measurement point and the overall population intercept for FB will be an estimate for average FB performance at Time 2.

At Level 1, the model was as follows:

$$FB_{ij} = \beta_{0i} + \beta_{1i}(\text{Time}) + e_{ij}.^{4}$$

[4] $FB_{ij}$ is the observed score of FB understanding at Time j for the $i^{th}$ child. $\beta_{1i}$ is the growth rate for the $i^{th}$ child over the data collection period and represents the expected change during a fixed unit of time. The intercept $\beta_{0i}$ is the initial status of FB understanding. $e_{ij}$ is the error which is independently and normally distributed with a mean of 0 and constant variance, $\sigma^2$.

Table 6
*Hierarchical Regression Analyses Predicting False Belief From Nonverbal IQ, Short-Term Memory, Inhibition, Reynell Language Scores, and False Complements, Controlling for Age and Initial False Belief Performance in Study 2*

| Variable | Unstandardized β | *SE* β | Standardized β | $R^2/\Delta R^2$ |
|---|---|---|---|---|
| Time 1 to Time 2 ($n$ = 72) | | | | |
| Step 1 | | | | .25** |
| Age | 1.13 | .58 | .26 | |
| False belief, Time 1 | 0.48 | .18 | .29** | |
| Nonverbal IQ | 0.01 | .11 | .01 | |
| Short-term memory | 0.39 | .21 | .23 | |
| Inhibition–Total | −0.02 | .21 | −.01 | |
| Step 2 | | | | .06* |
| Reynell–Total | 0.08 | .03 | .39* | |
| Step 3 | | | | .01 |
| False complements, Time 1 | 0.18 | .24 | .09 | |
| Time 1 to Time 3 ($n$ = 70) | | | | |
| Step 1 | | | | .30**** |
| Age | 1.03 | .59 | .23 | |
| False belief, Time 1 | 0.58 | .18 | .34** | |
| Nonverbal IQ | 0.10 | .11 | .12 | |
| Short-term memory | 0.22 | .22 | .12 | |
| Inhibition–Total | 0.17 | .21 | .10 | |
| Step 2 | | | | .18**** |
| Reynell–Total | 0.15 | .03 | .72**** | |
| Step 3 | | | | .00 |
| False complements, Time 1 | 0.09 | .22 | .04 | |
| Time 2 to Time 3 ($n$ = 70) | | | | |
| Step 1 | | | | .58**** |
| Age | 0.13 | .46 | .03 | |
| False belief, Time 2 | 0.72 | .09 | .69**** | |
| Nonverbal IQ | 0.09 | .08 | .11 | |
| Short-term memory | 0.01 | .17 | .00 | |
| Inhibition–Total | 0.17 | .16 | .10 | |
| Step 2 | | | | .08**** |
| Reynell–Total | 0.11 | .03 | .50**** | |
| Step 3 | | | | .01 |
| False complements, Time 2 | 0.28 | .17 | .14 | |
| Time 1 to Time 4 ($n$ = 71) | | | | |
| Step 1 | | | | .34**** |
| Age | 0.83 | .55 | .19 | |
| False belief, Time 1 | 0.60 | .17 | .37** | |
| Nonverbal IQ | −0.08 | .10 | −.09 | |
| Short-term memory | 0.59 | .20 | .35** | |
| Inhibition–Total | 0.13 | .20 | .08 | |
| Step 2 | | | | .14**** |
| Reynell–Total | 0.13 | .03 | .62**** | |
| Step 3 | | | | .00 |
| False complements, Time 1 | −0.11 | .21 | −.05 | |
| Time 2 to Time 4 ($n$ = 71) | | | | |
| Step 1 | | | | .45**** |
| Age | 0.16 | .51 | .04 | |
| False belief, Time 2 | 0.54 | .10 | .53**** | |
| Nonverbal IQ | −0.07 | .09 | −.09 | |
| Short-term memory | 0.43 | .19 | .25* | |
| Inhibition–Total | 0.10 | .18 | .06 | |
| Step 2 | | | | .06** |
| Reynell–Total | 0.09 | .03 | .44* | |
| Step 3 | | | | .01 |
| False complements, Time 2 | 0.23 | .20 | .12 | |

Table 6 (*continued*)

| Variable | Unstandardized β | *SE* β | Standardized β | $R^2/\Delta R^2$ |
|---|---|---|---|---|
| | | Time 3 to Time 4 ($n$ = 70) | | |
| Step 1 | | | | .75**** |
| Age | −0.01 | .35 | −.00 | |
| False belief, Time 3 | 0.80 | .07 | .82**** | |
| Nonverbal IQ | −0.15 | .06 | −.19 | |
| Short-term memory | 0.44 | .13 | .26** | |
| Inhibition–Total | −0.03 | .12 | −.02 | |
| Step 2 | | | | .00 |
| Reynell–Total | 0.02 | .03 | .08 | |
| Step 3 | | | | .01 |
| False Complements, Time 3 | −0.27 | .15 | −.16 | |

* $p < .05$. ** $p < .01$. **** $p < .001$.

At Level 2, the model was as follows:

$$\beta_{0i} = \pi_{00} + \gamma_{0i}$$

$$\beta_{1i} = \pi_{10} + \gamma_{1i}.^5$$

The results showed that the intercept, $\pi_{00}$, was significantly different from zero, $t(68) = 8.56$, $p < .001$, indicating that, on average, children's FB scores at Time 2 were not at floor. In addition, the slope, $\pi_{10}$, was also significantly different from zero, $t(68) = 4.79$, $p < .001$, indicating that, on average, children' FB understanding improved significantly from Time 2 to Time 4. Moreover, significant variability existed in both the random intercept, $\beta_{0i}$, $T_0^2 = 5.72$, $p < .001$, and the random slope, $\beta_{1i}$, $T_1^2 = 0.80$, $p < .001$, indicating that it would be appropriate to explore the role of other variables which might contribute to the growth and variability in FB understanding over this time period.

> *Hypothesis 2:* Individual differences in FB intercepts and trajectories can be captured by individual differences in false complements.

In order to test the simplest version of the "complements as precursors" to FB model, we chose to enter children's initial differences in Complement performance (centered on the grand mean) at Time 1 as our primary individual differences variable at Level 2 of the model. This model was identical to the previous model at Level 1, and at Level 2 it was as follows:

$$\beta_{0i} = \pi_{00} + \pi_{01}\,(\text{False Complements at T1}) + \gamma_{0i}$$

$$\beta_{1i} = \pi_{10} + \pi_{11}\,(\text{False Complements at T1}) + \gamma_{1i}.$$

These results showed that children's performance on false complements at Time 1 was a significant predictor of that intercept, $t(67) = 2.60$, $p < .01$. However, in this model, children's initial performance on false complements did not significantly contribute to the growth (slope) of FB over time, $t(67) = -0.87$, $p = .39$. Moreover, significant variability continued to exist in both the intercept, $\beta_{0i}$, $T_0^2 = 5.19$, $p < .001$, and slope, $\beta_{1i}$, $T_1^2 = 0.80$, $p < .001$. Thus, because individual differences in false complements at Time 1 predicted variation in FB performance at Time 2 (intercept measure, $\beta_0$), we included false complements in our further models that tried to examine the role of individual differences in language and executive functioning on FB understanding and to control for individual differences in children's age and initial levels of FB understanding at Time 1.

> *Hypothesis 3:* Linear growth of FB can be accounted for by initial state of complements even when initial FB performance and age are taken into account.

Next, given that children's initial FB performance, although low, was not at zero and that age, initial complements, and FBs are so highly correlated, we tested a model that also included age and initial FB performance as Level 2, or individual differences, variables. For this hypothesis, the Level 2 model centered the child's age, false complement, and initial FB score at Time 1 at each of the respective grand means.

As would be expected from the regression analyses, the predictability of false complements in this model decreased appreciably. Nonetheless, both initial complement and initial FB performance remained marginally significant for the intercept, $\beta_{0i}$, $t(65) = 1.92$ and $t(65) = 1.64$, $ps < .10$, respectively. In this model, age was a significant predictor of the intercept, $\beta_{0i}$, $t(65) = 3.49$, $p < .001$, but none of these factors significantly predicted variation in the slope of FB performance.[6] It should be noted that although variability for the random intercept decreased in this model, $\beta_{0i}$, $T_0^2 = 4.22$, $p < .001$, the variability of the random slope, $\beta_{1i}$, did not

[5] $\pi_{00}$ is the mean initial status of FB understanding. $\pi_{10}$ is the mean growth rate. $\gamma_{0i}$ and $\gamma_{1i}$ represent the unique random variation associated with person i.

[6] Note that one could argue that controlling for initial FB performance is inappropriate given the short time intervals between testing and that the correlation between initial FB and later FB performance is moderately high. Although we feel that it is statistically more appropriate to control for initial FB performance, given that performance was not at zero, this and all of our subsequent models were run in both ways. For Hypothesis 3, the effect of false complement performance on the intercept, $\beta_0$, goes from marginal to significant, $t(66) = 2.46$, $p < .05$, with no differences in the other parameters of the model, when FB performance at Time 1 is no longer controlled.

Table 7
*HLM Components Predicting False Belief Performance in Study 2*

| Hypothesis 4A | Fixed effects Coefficient | *SE* | *t*(61) | Random effects Variance |
|---|---|---|---|---|
| Intercept ($\beta_0$) | | | | 3.75**** |
| Intercept $\pi_{00}$ | 2.70 | 0.26 | 10.57**** | |
| Complement at Time 1(T1), $\pi_{01}$ | 0.31 | 0.22 | 1.40 | |
| Age at T1, $\pi_{02}$ | 0.69 | 0.51 | 1.35 | |
| False belief at T1, $\pi_{03}$ | 0.32 | 0.16 | 1.97* | |
| Nonverbal IQ at Time 2 (T2), $\pi_{04}$ | −0.02 | 0.10 | −0.17 | |
| Short-term memory at T2, $\pi_{05}$ | −0.03 | 0.23 | −0.13 | |
| Inhibition at T2, $\pi_{06}$ | −0.16 | 0.26 | −0.61 | |
| Reynell language at T2, $\pi_{07}$ | 0.09 | 0.03 | 2.85*** | |
| Slope ($\beta_1$) | | | | 0.80**** |
| Intercept $\pi_{10}$ | 0.68 | 0.14 | 4.89**** | |
| Complement at T1, $\pi_{11}$ | −0.16 | 0.10 | −1.53 | |
| Age at T1, $\pi_{12}$ | −0.34 | 0.26 | −1.33 | |
| False belief at T1, $\pi_{13}$ | 0.10 | 0.09 | 1.07 | |
| Nonverbal IQ at T2, $\pi_{14}$ | −0.06 | 0.05 | −1.13 | |
| Short-term memory at T2, $\pi_{15}$ | 0.04 | 0.13 | 0.29 | |
| Inhibition at T2, $\pi_{16}$ | 0.03 | 0.12 | 0.23 | |
| Reynell language at T2, $\pi_{17}$ | 0.03 | 0.02 | 1.53 | |

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .005$. $^{****}p < .001$.

change, $T_1^2 = 0.82$, $p < .001$. Thus, it is important to consider what other factors might account for the relationship between FB and false complement performance and, more generally, for the growth of FB understanding in this sample.

> *Hypothesis 4A:* Linear growth of FB can be accounted for by individual variation in language, executive function, or nonverbal intelligence even when initial FB performance and age are taken into account.

In this model, we included all of the variables from the previous model, with all Level 2 variables centered at their grand means, as well as each of our control measures (either centered at their grand means or *z*-scores, which were already centered) such that the Level 2 model was as follows:

$\beta_{0i} = \pi_{00} + \pi_{01}$ (false complements at T1) $+ \pi_{02}$ (age at T1) $+ \pi_{03}$ (FB at T1) $+ \pi_{04}$ (nonverbal IQ at T2) $+ \pi_{05}$ (short-term memory at T2) $+ \pi_{06}$ (inhibition at T2) $+ \pi_{07}$ (Reynell language at T2) $+ \gamma_{0i}$

$\beta_{1i} = \pi_{10} + \pi_{11}$ (false complements at T1) $+ \pi_{12}$ (age at T1) $+ \pi_{13}$ (FB at T1) $+ \pi_{14}$ (nonverbal IQ at T2) $+ \pi_{15}$ (short-term memory at T2) $+ \pi_{16}$ (inhibition at T2) $+ \pi_{17}$ (Reynell language at T2) $+ \gamma_{1i}$.

As can be seen in Table 7, only FB performance at Time 1 and the Reynell measure of children's language abilities at Time 2 significantly contributed to predicting the children's FB performance at Time 2.[7] However, none of these factors accounted for the change in FB performance from Time 2 to Time 4. To ensure that this was a real relationship and not an artifact of having a large number of intercorrelated variables entered into the model, we also tried several more parsimonious versions of this model, including several models that did not include the language measures and models that aggregated the short-term memory and inhibition scores, but none of these other models resulted in significant predictions of either the intercept or slope for FB performance by any other of the Time 2 measures (nonverbal IQ, short-term memory, inhibition). Adding gender to the model, however, did alter the relationships slightly. Specifically, although gender did not significantly predict the intercept for FB at Time 2, it did significantly predict the change in FB performance from Time 2 to Time 4, $\beta_1$, $t(62) = -1.97$, $p < .05$, indicating that the rate of change in FB performance for girls was higher than for boys, when all other variables were controlled. Importantly, however, adding gender to the model did not improve the predictability of false complements or any of the control measures for either the intercept or slope.

Thus, in essence both echoing and going beyond the traditional linear regression models, the HLM analyses suggested a strong, consistent effect of general language abilities, and only these abilities, on FB performance in our study. As in the traditional regression analyses, we did not find evidence for a strong relationship between complements and FB once we controlled for age and initial levels of FB performance. Somewhat to our surprise, however, we also did not find that short-term memory or any of the other control variables contributed significantly to either the intercept or slope of FB performance in our sample of Cantonese-speaking children.

[7] For Hypothesis 4, the effect of false complement performance on the intercept, $\beta_0$, goes from nonsignificant to marginal, $t(63) = 1.90$, $p < .10$, with no differences in the other parameters of the model, when FB performance at Time 1 is no longer controlled.

However, we were puzzled by two issues in these data. The first is that false complements were hovering at the brink of significance in several models, with marginal or close to marginal *p* values, particularly when the effects of initial FB performance were not controlled. The second is that the similar appearance of short-term memory effects in the regression analyses and its correlation with all of the relevant measures led us to wonder whether short-term memory might not still be an important factor in understanding why there might be a relationship between Cantonese-speaking children's complement understanding and FB performance in both Study 1 and Study 2. To pursue these concerns, we decided to examine whether any of the control measures (or FB) would be predictive of false complement performance over the three testing points. If FB was as (or more) predictive of complements than complements of FB, then this would definitively argue against the complements as precursors hypothesis. Also, if short-term memory or other control measures significantly predicted complement performance, then this might explain why false complements hovered near significance in our HLM analyses and short-term memory measures emerged as significant, although nonunique, predictors in the regression analyses.

Although we ran a full set of HLM analyses paralleling those for FB performance (but with false complements as the dependent variable), we report only two of these (see Table 8).

> *Hypothesis 1B:* Linear growth of false complements over time.

As with the growth model for FB, the best fitting growth model for performance on false complements was linear, with no additional variance being explained by the quadratic model. As with FB, false complements showed a significant intercept, $\pi_{00}$, $t(68) = 8.99$, $p < .001$, indicating that, on average, children's false complement scores at Time 2 were not at floor. However, unlike FB, the slope, $\pi_{10}$, for complements did not reach significance, $t(68) = 1.19$, $p > .10$, indicating that, on average, children' false complement understanding did not improve significantly from Time 2 to 4. Nonetheless, significant variability existed in both the random intercept, $\beta_{0i}$, $T_0^2 = 1.01$, $p < .001$, and the random slope, $\beta_{1i}$, $T_1^2 = 0.19$, $p < .05$, indicating that it would be appropriate to explore the role of other variables which might contribute to the variability in complement understanding both at Time 2 and the changes between Times 2 and 4 for this sample.

> *Hypothesis 4B:* Linear growth of false complements can be accounted for by individual variation in language, executive function, FB, or nonverbal intelligence even when initial false complement performance and age are taken into account.

In Table 8, we present the results for the full set of predictors of FC, which is parallel to the model generated for Hypothesis 4A, with FB as an outcome. Unlike the models for FB, neither false complement nor FB performance at Time 1 predicted false complement performance at Time 2 (intercept). Furthermore, short-term memory performance, but not the Reynell, was predictive of false complement performance. Moreover, although the overall means did not change significantly from Time 2 to Time 4 for false complement, individual differences in the rate of change were predicted by false complement performance at Time 1 and by both the Reynell and short-term memory at Time 2. This last relationship was negative (i.e., the slope for children with higher short-term memory at Time 2 was not as steep as the slope for children

Table 8
*HLM Components Predicting False Complement Performance in Study 2*

| Hypothesis 4B | Fixed effects | | | Random effects |
|---|---|---|---|---|
| | Coefficient | *SE* | $t(61)$ | Variance |
| Intercept ($\beta_0$) | | | | $0.75^{****}$ |
| Intercept $\pi_{00}$ | 1.43 | 0.15 | $9.72^{****}$ | |
| Complement at Time 1 (T1), $\pi_{01}$ | −0.07 | 0.13 | −0.52 | |
| Age at T1, $\pi_{02}$ | 0.78 | 0.32 | $2.46^{*}$ | |
| False Belief at T1, $\pi_{03}$ | 0.04 | 0.10 | 0.42 | |
| Nonverbal IQ at Time 2 (T2), $\pi_{04}$ | −0.07 | 0.05 | −1.23 | |
| Short-term Memory at T2, $\pi_{05}$ | 0.29 | 0.12 | $2.34^{*}$ | |
| Inhibition at T2, $\pi_{06}$ | −0.09 | 0.12 | −0.76 | |
| Reynell Language at T2, $\pi_{07}$ | 0.00 | 0.02 | 0.01 | |
| Slope ($\beta_1$) | | | | 0.05 |
| Intercept $\pi_{10}$ | 0.11 | 0.08 | 1.27 | |
| Complement at T1, $\pi_{11}$ | 0.13 | 0.07 | 1.82† | |
| Age at T1, $\pi_{12}$ | −0.04 | 0.18 | −0.23 | |
| False Belief at T1, $\pi_{13}$ | 0.08 | 0.06 | 1.49 | |
| Nonverbal IQ at T2, $\pi_{14}$ | −0.01 | 0.03 | −0.32 | |
| Short-term Memory at T2, $\pi_{15}$ | −0.25 | 0.07 | $-3.54^{***}$ | |
| Inhibition at T2, $\pi_{16}$ | 0.08 | 0.07 | 1.21 | |
| Reynell language at T2, $\pi_{17}$ | 0.03 | 0.01 | $2.60^{**}$ | |

$^{*}p < .05$. $^{**}p < .01$. $^{***}p < .005$. $^{****}p < .001$.

with lower short-term memory scores), however, and may account for part of the inconsistencies in the previous analyses. Nonetheless, it is important to note that false complement performance was not predicted by FB performance and that both short-term memory and general language competence emerged as significant predictors of both false complement and FB.

Thus, given that the overt syntax of belief does not differ from desire and that false complements are not required for the expression of FB statements in Cantonese (the lexical term *ji5wai4* serves this purpose instead), it is more likely that the correlation between FB and false complement is a function of the combined effects of having an adequate short-term memory to hold the various sentence components in mind as well as general language skills.

## General Discussion

Both of the studies (one cross-sectional, one longitudinal) reported in this article initially found a positive relationship between children's facility with a specific syntactic structure (false complements) and FB understanding in Cantonese-speaking children. However, in our studies, as in other studies examining the nature of this relationship longitudinally, once initial levels of FB performance and age were controlled, false complements no longer significantly and uniquely predicted FB. In contrast, general language abilities were consistent predictors of FB understanding, even after controlling for false complements, initial FB, age, short-term memory, inhibition, and nonverbal IQ. In addition, general language abilities and short-term memory were found to be predictive of false complement performance. This suggests that, for Cantonese, complements are not necessary precursors to an understanding of FB, which is what one would expect given that the finite complement syntax central to de Villier's (1995) hypothesis does not distinguish belief from desire in Cantonese. Nonetheless, the fact that we (and Cheung et al., 2004) found any correlations between performance on such a task, and that studies with English-speaking children found evidence for perhaps stronger relations between complements and FB understanding, returns us to the more general question of how and why children's language competence is related to their performance on FB tasks.

To answer this more general question, let us consider this hypothesis in the light of other hypotheses (and other evidence) about the nature of the relationship between language and FB. Specifically, it is important to draw a parallel between de Villiers's (1995) hypothesis about false complements as a specific syntactic structure that may foster FB understanding and other hypotheses that have suggested specific lexical terms as facilitative of or prerequisite to FB. In each of these hypotheses, there is an attempt to go beyond the general link between language and FB to the search for a specific mechanism or linguistic device that might be responsible for this link.

As outlined in the introduction, some authors (e.g., Olson, 1988) have suggested that there is a need for children to first develop metacognitive terms before they are able to understand FB. Although Moore et al. (1990) found initial support for this hypothesis, their study was not longitudinal, nor did it control for children's overall language abilities or initial levels of FB understanding. Thus, it is not clear whether knowledge of metacognitive terms would uniquely predict FB performance when these other factors were controlled for. In studies that directly compared syntactic versus semantic aspects of language development (e.g., Astington & Jenkins, 1999; de Villiers & Pyers, 2002), syntax, but not semantics, was found to predict FB in English, thus suggesting that metacognitive vocabulary may not play a unique role.

In cross-linguistic studies that examined the effects of specific lexical terms that mark beliefs as false (vs. neutral), the use of these FB terms was shown to have a facilitating effect. However, this too was limited. Specifically, these terms only helped when they were used in the test question and only with children who were already at the cusp of understanding FB (i.e., 4- and 5-year-olds, not 3-year-olds). Use of the lexically marked terms never improved performance beyond a nonmental state verb version of the task, nor did they result in better performance than that shown by children who spoke languages which did not mark FB lexically (Shatz et al., 2003; Tardif et al., 2004).

Taken together, and with the results from studies with deaf children, we believe these studies show that there is a special relationship between language and FB development. What they do not support, however, is the conclusion that any particular linguistic device (semantic or syntactic) that marks FB has a prerequisite effect on the development of FB understanding. This is true both cross-linguistically (English-speaking children do not have lexical terms specific to FB, and yet they acquire the concept; Chinese-speaking children do not have distinctions between finite and nonfinite complements that correspond with the belief-desire distinction, and yet they also acquire the concept) and within a language as well. For instance, children in Lohmann and Tomasello's (2003) study improved on FB performance even when mental state terms and finite complements were specifically avoided in discourse about deceptive objects.

The data on whether or not specific linguistic devices facilitate FB understanding, however, is more complex. Importantly, all of the languages discussed so far (English, Cantonese, German, Mandarin, Spanish, Portuguese) mark FB differently—English marks it with complement syntax, Mandarin and Cantonese with lexical terms, and Cantonese and other languages can also mark FB with sentence-final evidentiality particles (Cheung et al., 2004). Nonetheless, FB understanding is accomplished in all of these languages, regardless of the specific marker or device that is used to mark FB. And yet studies with English suggest that training on the ways in which one's language marks FB do facilitate FB performance (Hale & Tager-Flusberg, 2003; Lohmann & Tomasello, 2003) but that these training effects are also limited. Children must be at the cusp of understanding FB, and training on false complements only is no better than discourse that avoids finite complements but focuses on the deceptive qualities of objects (Lohmann & Tomasello, 2003). Similarly, studies with other languages suggest that FB understanding is facilitated when parents or experimenters use the language-specific devices for talking about FBs but, again, only when children are at the cusp of this understanding. In our Study 2, individual differences in children's language abilities also consistently predicted FB performance at Time 2. Overall, then, these data suggest that language is an important facilitator for FB understanding and that children are remarkably good at cuing in to whatever means their language uses to mark FB to help them achieve this understanding, but they do so at remarkably similar ages, regardless of the language spoken or of the device that is used. This, then, leads us to wonder what other factors might be responsible for this relationship.

Unlike in previous studies conducted with English (e.g., Carlson et al., 2002, 2004; Hughes, 1998b), the inhibitory control measures did not emerge as unique predictors of FB performance for our sample of Cantonese-speaking children. However, short-term memory did emerge as a significant predictor in our regression models before general language abilities were controlled, and it also emerged, together with general language abilities, as a significant predictor of false complement performance in our HLM analyses. Although there may be a number of reasons for this discrepancy (e.g., differences in the age, level of FB performance, time periods, or measures used to examine the data), it is also the case that these abilities are all intercorrelated in development such that different children may rely more heavily on one or the other of these abilities. Specifically, although executive functioning abilities may show strong relationships to FB in an English-speaking sample of preschoolers, they might not show the same relationships in a sample of children for whom these abilities might be developed at different times or through different pathways. An important suggestion in this direction is offered by comparing performance on our sample's control versus conflict trials in the Luria hand game to that shown by the preschoolers in the Hughes (1998b) study. Of interest is that the mean number of trials to criterion on this task in the control trials were almost identical in the two studies ($M = 6.31$, $SD = 1.38$ for our study vs. $M = 6.8$, $SD = 2.3$ for Hughes' 1998b, study), whereas performance on the conflict conditions was very different ($M = 7.49$, $SD = 3.22$ for our study vs. $M = 10.4$, $SD = 4.2$ for Hughes, 1998b). It is the conflict condition that is of special interest in measures of executive functioning, and this difference, in the light of similarities with the control condition, suggests that there may indeed be differences in the rate with which Hong Kong and British children, at least, develop the ability to inhibit a prepotent response. Additionally, a similar difference in performance in executive functioning (but not FB) tasks between Mandarin-speaking Chinese children and English-speaking American children was observed in a recent study (Sabbagh, Xu, Carlson, Moses, & Lee, 2006) with a much larger battery of EF tasks. However, in that study, a relationship between EF and FB was found within the sample even though the consistently superior performance on EF was not paralleled by superior performance on FB tasks when looking across cultures. Again, this could be due to differences in method (measures, cross-sectional vs. longitudinal designs), or it could be due to differences in paths toward achieving FB. Although we prefer the latter explanation, there is clearly much room for future exploration of this issue in longitudinal samples both within and across languages.

Our tentative conclusion, therefore, is that just as there may be language-specific markers of FB, there may also be language- or culturally specific associations between FB and other factors that are developing within the preschool years. Investigating the relationships between these different emerging abilities is fundamental to furthering our understanding of what mechanisms may or may not underlie those developments and also of the flexibility inherent in each of these systems. However, to investigate these relationships more thoroughly requires a depth and a breadth of data on individual children across multiple time points in different cultures. Moreover, it requires sophisticated analytical tools such as HLM techniques, which can allow us to examine both population-level and individual differences effects in the trajectories of simultaneously developing skills. Neither of these approaches were readily available a decade ago, and today, the data are still costly and difficult to obtain. But perhaps an analogy to the effects of language-specific markers and children's developing FB understanding can urge us to consider them more seriously—they are tools that are not prerequisite to further our understanding, but they will most certainly facilitate that understanding.

## References


Astington, J. W. (2000). *Minds in the making: Essays in honor of David R. Olson.* Malden, MA: Blackwell.

Astington, J. W., & Jenkins, J. M. (1999). A longitudinal study of the relation between language and theory-of-mind development. *Developmental Psychology, 35,* 1311–1320.

Bartsch, K., & Wellman, H. M. (1995). *Children talk about the mind.* New York: Oxford University Press.

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101,* 147–158.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Thousand Oaks, CA: Sage.

Carlson, S. M., Mandell, D. J., & Williams, L. (2004). Executive function and theory of mind: Stability and prediction from ages 2 to 3. *Developmental Psychology, 40,* 1105–1122.

Carlson, S. M., Moses, L. J., & Breton, C. (2002). How specific is the relation between executive function and theory of mind? Contributions of inhibitory control and working memory. *Infant & Child Development, 11,* 73–92.

Cassidy, K. W. (1998). Preschoolers' use of desires to solve theory of mind problems in a pretense context. *Developmental Psychology, 34,* 503–511.

Cheung, H., Chen, H.-C., Creed, N., Ng, L., Wang, S. P., & Mo, L. (2004). Relative roles of general and complementation language in theory-of-mind development: Evidence from Cantonese and English. *Child Development, 75,* 1155–1170.

Cutting, A. L., & Dunn, J. (1999). Theory of mind, emotion understanding, language, and family background: Individual differences and interrelations. *Child Development, 70,* 853–865.

de Villiers, J. (1995, April). *Steps in the mastery of sentence complements.* Paper presented at the biennial meeting of the Society for Research in Child Development. Indianapolis, IN.

de Villiers, J., & de Villiers, P. A. (2000). Linguistic determinism and the understanding of false belief. In P. Mitchell & K. J. Riggs. (Eds.), *Children's reasoning and the mind* (191–228). Hove, England: Taylor & Francis.

de Villiers, J., & Pyers, J. E. (2002). Complements to cognition: A longitudinal study of the relationship between complex syntax and false-belief-understanding. *Cognitive Development, 17,* 1037–1060.

Diamond, A., & Taylor, C. (1996). Development of an aspect of executive control: Development of the abilities to remember what I said and to "Do as I say, not as I do." *Developmental Psychobiology, 24,* 315–334.

Flavell, J. H. (2004). Theory-of-mind development: Retrospect and prospect. *Merrill-Palmer Quarterly, 50,* 274–290.

Flynn, E., O'Malley, C., & Wood, D. (2004). A longitudinal, microgenetic study of the emergence of false belief understanding and inhibition skills. *Developmental Science, 7,* 103–115.

Gale, E., de Villiers, P., de Villiers, J., & Pyers, J. (1996). Language and theory of mind in oral deaf children. In A. Stringfellow, D. Cahana-

Amitay, E. Hughes, & A. Zukowski (Eds.), *Proceedings of the Boston University Conference on Language Development* (Vol. 1, pp. 213–224). Somerville, MA: Cascadilla Press.

Gerstadt, C. L., Hong, Y. J., & Diamond, A. (1994). The relationship between cognition and action: Performance of children 3 1/2–7 years old on a Stroop-like day–night test. *Cognition, 53,* 129–153.

Goldman, R., Fristoe, M., & Woodcock, R. W. (1974). *Goldman-Fristoe-Woodcock Auditory Memory Tests.* Circle Mines, MN: American Guidance Service.

Gopnik, A., & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 257–293). New York: Cambridge University Press.

Hale, C. M., & Tager-Flusberg, H. (2003). The influence of language on theory of mind: A training study. *Developmental Science, 6,* 346–359.

Happé, F. G. E. (1995). The role of age and verbal ability in the theory of mind task performance of subjects with autism. *Child Development, 66,* 843–855.

Hogrefe, G.-J., Wimmer, H., & Perner, J. (1986). Ignorance verses false belief: A developmental lag in attribution of epistemic states. *Child Development, 57,* 567–582.

Hughes, C. (1998a). Executive function in preschoolers: Links with theory of mind and verbal ability. *British Journal of Developmental Psychology, 16,* 233–253.

Hughes, C. (1998b). Finding your marbles: Does preschoolers' strategic behavior predict later understanding of mind? *Developmental Psychology, 34,* 1326–1339.

Hughes, C., & Dunn, D. (1997). "Pretend you didn't know": Preschoolers' talk about mental states in pretend play. *Cognitive Development, 12,* 381–403.

Hughes, C., Dunn, J., & White, A. (1998). Trick or treat? Uneven understanding of mind and emotion and executive dysfunction in "hard-to-manage" preschoolers. *Journal of Child Psychology and Psychiatry, 39,* 981–994.

Jenkins, J. M., & Astington, J. W. (1996). Cognitive factors and family structure associated with theory of mind development in young children. *Developmental Psychology, 32,* 70–78.

Kloo, D., & Perner, J. (2003). Training transfer between card sorting and false belief understanding: Helping children apply conflicting descriptions. *Child Development, 74,* 1823–1839.

Lee, K., Olson, D. R., & Torrance, N. (1999). Chinese children's understanding of false beliefs: The role of language. *Journal of Child Language, 26,* 1–21.

Lillard, A. (1998). Ethnopsychologies: Cultural variations in theories of mind. *Psychological Bulletin, 123,* 3–32.

Liu, D., Wellman, H. M., Tardif, T., & Sabbagh, M. A. (2006). *Theory of mind development in China: Meta-analysis of Chinese children's understanding of false belief.* Unpublished manuscript.

Lohmann, H., & Tomasello, M. (2003). The role of language in the development of false belief understanding: A training study. *Child Development, 74,* 1130–1144.

Luria, A. R. (1965). Two kinds of motor perseveration in massive injury of the frontal lobes. *Brain, 88,* 1–10.

Mitchell, P. (1994). Realism and early conception of mind: A synthesis of phylogenetic and ontogenetic issues. In C. Lewis & P. Mitchell (Eds.), *Children's early understanding of mind* (pp. 19–46). Mahwah, NJ: Erlbaum.

Moore, C., Bryant, D., & Furrow, D. (1989). Mental terms and the development of certainty. *Child Development, 60,* 167–171.

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,* 722–730.

Olson, D. R. (1988). On the origins of beliefs and other intentional states in children. In J. W. Astington, P. L. Harris, & D. R. Olson (Eds.), *Developing theories of mind* (pp. 414–426). New York: Cambridge University Press.

Perner, J. (1991). *Understanding the representational mind.* Cambridge, MA: MIT Press.

Perner, J., Sprung, M., & Zauner, P. (2003). Want that is understood well before say that, think that, and false belief: A test of de Villier's linguistic determinism on German-speaking children. *Child Development, 74,* 179–188.

Peterson, C. C., & Siegal, M. (1995). Deafness, conversation and theory of mind. *Journal of Child Psychology and Psychiatry, 36,* 459–474.

Reynell, J. & Huntley, M. (1985). *Reynell Developmental Language Scales* (2nd ed.) London: NFER-Nelson.

Roeper, T., & de Villiers, J. (1994). Lexical links in the WH-chain. In B. Lust, G. Hermon, & J. Kornfilt (Eds.), *Syntactic theory and first language acquisition: Cross-linguistic perspectives, Vol. 2: Binding, dependencies, and learnability* (pp. 357–390). Mahwah, NJ: Erlbaum.

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 U.S. preschoolers. *Psychological Science, 17,* 74–81.

Schulenberg, J., & Maggs, J. L. (2001). Moving targets: Modeling developmental trajectories of adolescent alcohol misuse, individual and peer risk factors, and intervention effects. *Applied Developmental Science, 5,* 237–253.

Shatz, M., Diesendruck, G., & Martinez-Beck, I. (2003). The influence of language and socioeconomic status on children's understanding of false belief. *Developmental Psychology, 39,* 717–729.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics.* New York: HarperCollins.

Tager-Flusberg, H. (1993). What language reveals about the understanding of minds in children with autism. In S. Baron-Cohen, H. Tager-Flusberg, & D. Cohen (Eds.), *Understanding other minds: Perspectives from autism* (pp. 138–157). Oxford, England: Oxford University Press.

Tardif, T., & Wellman, H. M. (2000). Acquisition of mental state language in Mandarin- and Cantonese-speaking children. *Developmental Psychology, 36,* 25–43.

Tardif, T., Wellman, H. M., & Cheung, K. M. (2004). False belief understanding in Cantonese-speaking children. *Journal of Child Language, 31,* 779–800.

Thorndike, R. L., Hagan, E. P., & Sattler, J. M. (1986). *Stanford–Binet Intelligence Scales* (4th ed.). Itasca, IL: Riverside Publishing.

Vinden, P. G. (1996). Junin Quechua children's understanding of mind. *Child Development, 67,* 1707–1716.

Vinden, P. G. (1999). Children's understanding of mind and emotion: A multi-culture study. *Cognition and Emotion, 13,* 19–48.

Vinden, P. G. (2001). Parenting attitudes and children's understanding of mind: A comparison of Korean–American and Anglo-American families. *Cognitive Development, 16,* 793–809.

Vinden, P. G. (2002). Understanding minds and evidence for belief: A study of Mofu children in Cameroon. *International Journal of Behavioral Development, 26,* 445–452.

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.

Wellman, H. M., & Gelman, S. A. (1992). Cognitive development: Foundational theories of core domains. *Annual Review of Psychology, 43,* 337–375.

Wellman, H. M., & Liu, D. (2004). Scaling of theory-of-mind tasks. *Child Development, 75,* 523–541.

Wimmer, H., & Perner, J. (1983). Beliefs about beliefs: Representation and constraining function of wrong beliefs in young children's understanding of deception. *Cognition, 13,* 103–128.

## Appendix

### Complement Comprehension Task for Study 1 and Time 1 and Time 3 in Study 2

1. 阿爸 話俾 男仔 聽 佢 食緊 隻蛋，
Aa3baa4 waa6 bei2 naam4zai2 teng1 keoi5 sik6gan2 zek3dan2,
Dad tell boy listen he eating an egg,
但 其實 係 一個 乒乓波。
dan6 kei4sat6 hai6 jat1go3 ping1pong1 bo1
but really is one (classifier) ping-pong ball
"Dad told the boy that he was eating an egg, but it really was a ping-pong ball."

2. 老公 以為 佢 係 間房 見到 鬼，
lou5gung1 ji5wai4 keoi5 hai6 gan1 fong4 gin3 dou2 gwai2,
Dad think he in the room see ghost,
但 其實 係 一張 氈。
dan6 kei4sat6 hai6 jat1 zoeng1 zin1
but really is one (classifier) blanket
"Dad thought he saw a ghost in the room, but it really was a blanket."

3. 先生 話俾 個 女仔 聽
Sin1saang1 waa6 bei2 go3 neoi5zai2 teng1
The teacher tell this girl listen
有隻 昆蟲 係 佢 頭髮度，
jau5zek3 kwan1cung4 hai2 keoi5 tau4faat3 dou6,
an insect in her hair,
但 其實 係 一塊 樹葉。
dan6 kei4sat6 hai6 jat1 faai3 syu6jip6
but really is one (classifier) leave
"The teacher told the girl there was an insect in her hair. but it really was a leaf."

4. 先生 以為 個 女仔 畫左 塊面，
Sin1saang1 ji5wai4 go3 neoi5zai2 waak6 zo2 faai3min6,
Teacher think this girl draw a face,
但 佢 其實 係 張紙 上面 亂咁畫。

False Belief and Complements in Cantonese

dan6 keoi5 kei4sat6 hai6 zoeng1zi2 soeng6min6 lyun6 gam3 waak6
But she really in paper on draw scribbles.
"The teacher thought the girl drew a face, but it really was a scribble.

5. 人地 話俾 佢 聽
jan4dei6 waa6 bei2 keoi5 teng1
Other tell him listen
佢 條褲 後面 穿左 個窿，
keoi5 tiu4 fu3 hau6min6 cyun1zo2 go3 lung1,
his pants behind cut a hole,

(*Appendix continues*)

但 其實 係 一張 紙。

dan6 kei4sat6 hai6 jat1zoeng1 zi2

but really is one piece of paper.

"Someone told him his pants had a hole, but it really was a piece of paper."

6. 人地 話俾 佢 聽

jan4dei6 waa6 bei2 keoi5 teng1

Other tell him listen

有隻 烏蠅 係 佢 碗 麥粥 入面，

jau5zek3 wo1jing1 hai6 keoi5 wun2 mak6zuk1 jap6min3,

a fly in his bowl cereal inside

但 其實 係 一粒 提子。

dan6 kei4sat6 hai6 jat1lap1 tai5zi2

But really is one (classifier) grape

"Someone told him there was a fly in his bowl of cereal, but it really was a grape."

7. 先生 以為 個 女仔 睇緊書，

Sin1saang1 ji5wai4 go3 neoi5zai2 tai2 gan2 syu1,

Teacher think this girl reading book,

但 佢 其實 玩緊 啤牌。

dan6 keoi5 kei4sat6 waan2gan2 pe1paai2

but she really playing cards

"The teacher thought the girl was reading book, but she was really playing cards."

8. 男仔 以為 阿爸 整損左 隻手，

False Belief and Complements in Cantonese

naam4zai2 ji5wai4 aa3baa4 zing2 syun2 zo2 zek3 sau2,

Boy think dad hurt hand,

但 其實 係 茄汁。

dan6 kei4sat6 hai6 ke2zap1

but really is ketchup

"He thought his dad had a cut, but it really was ketchup."

Received January 7, 2005
Revision received October 21, 2005
Accepted November 15, 2005 ■