How language does (and does not) relate to theory of mind: A longitudinal study of syntax, semantics, working memory and false belief

Lance Slade1,2* and Ted Ruffman2,3

1University of Surrey, Roehampton, UK
2University of Sussex, UK
3University of Otago, New Zealand

Forty-four children (mean 3.8 years) were given three false belief, a working memory, and four language tasks (each designed to tap a different aspect of syntax or semantics), and were tested again 6 months later. Once the range of scores in the language and false belief tasks were equated, there was a bidirectional relation between language and theory of mind. There was no evidence for syntax playing a unique role in the contribution of language to theory of mind. No one measure of syntax or semantics was more likely than any other to predict later false belief. Nor was false belief related more to one aspect of later language (syntax vs. semantics) than another. Our data, taken with other findings, are consistent with the idea that both syntax and semantics contribute to false belief understanding. Working memory did not mediate the relation between language and theory of mind, nor did it facilitate later false belief.

It is now generally accepted that children’s language ability is closely linked to their ability to pass false belief tasks, where false belief performance is taken as a critical marker of a child’s ‘theory of mind’ understanding. The precise nature of this relation has been a matter of some debate, however (e.g. see chapters in Lewis & Mitchell, 1994), and is still considered a central issue in current theory-of-mind research (Astington, 2001). Recently, two specific but related claims have been made about this relation. First, language ability might play a causal role in theory-of-mind development (i.e. language facilitates the later acquisition of a theory of mind but a theory of mind does not facilitate the later acquisition of language; Astington & Jenkins, 1999). Second, it has been argued that it is syntax that plays the significant role in theory-of-mind acquisition – though different aspects of syntax have been emphasized, from basic syntax such as word order (Astonigton & Jenkins, 1999) to more complex syntax such as...
relative clauses (Smith, Apperly, & White, 2003) or object complements (de Villiers & de

The link between children’s verbal ability and their ability to pass false belief tasks has
been demonstrated with a wide range of language tasks. Measures used as indices of
general verbal ability have included: (a) receptive vocabulary, where the child hears a
word and has to match it to the correct picture by pointing (e.g. Carlson & Moses, 2001;
Cutting & Dunn, 1999; Davies & Pratt, 1995; Doherty, 2000; Doherty & Perner, 1998;
Happe´, 1995; Hughes, 1998b; Perner, Lang, & Klo, 2002), (b) expressive narrative
speech, where the child has to retell a story from a sequence of pictures (e.g. Cutting &
Dunn, 1999; Hughes, 1998a), and (c) broader language tests tapping expressive and
receptive measures of syntax and semantics (e.g. Astington & Jenkins, 1999; Jenkins &

This basic association holds even after a number of potential confounds have been
considered. For example, since age correlates highly with language, it is possible that all
of the variance predicted by language can be accounted for by age. However
 correlations with language tasks, though reduced, remain significant even after age is
partialed out (e.g. Cutting & Dunn, 1999; Happé, 1995; Hughes, 1998a; Hughes &
Dunn, 1997; Jenkins & Astonig, 1996; Ruffman, Slade, & Crowe, 2002). Likewise, it
may be that correlations with language ability reflect the linguistic nature of false belief
tasks themselves, rather than the underlying conceptual understanding meant to be
tapped by the tasks (e.g. Chandler, Fritz, & Hala, 1989). Thus, Lewis and Osborne (1990)
and Siegal and Beattie (1991) both claimed that by clarifying the language of the test
question by including temporal markers (e.g. ‘where will the story character look first’),
even 3-year-olds do well on false belief tasks. Wellman, Cross, and Watson’s (2001) meta-
analysis specifically looked at a number of factors, such as temporal marking, that might
conceivably mask younger children’s competence. They showed that, although
temporal marking did improve performance, the effect was only significant for older
(4-year-old) children, but not 3-year-olds. Thus, while the claim that language ability has
an effect because of the linguistic demands of the task has some support, it still appears
that being able to pass false belief represents a genuine shift in conceptual
understanding (Perner, Leekam, & Leslie, 1987; but see Scholl & Leslie, 2001 for an
alternative viewpoint).

Crucial support that language ability plays an essential role in false belief
understanding comes from studies looking at the longitudinal relation between
language and false belief. Watson, Painter, and Bornstein (2001) and Farrar and Maag
(2002) showed that toddlers’ language ability at 2 years predicted false belief at 4 years
of age. Hughes (1998b) showed that children’s level of vocabulary at age 4 correlated
with theory of mind a year later. Ruffman, Slade, Rowlandson, Rumsey, and Garnham
(2003) showed that general language (tapping both syntax and semantics) at 3 years
predicted belief at 3 1/2, 4 and 5 1/2 years. Astington and Jenkins (1999) used a cross-lagged
design with 59 three-year-olds, giving them false belief tasks and a general language test
(the TELD, Test of Early Language Development; Hresko, Reid, & Hammill, 1981) at
three time points over 7 months. Language predicted theory of mind over two of the
three time points (explaining between 11% and 13% of the variance after controlling
earlier age and theory of mind), but theory of mind did not predict later language at any
time-point (after controlling earlier age and language). This is clearly consistent with
language ability causally facilitating later theory of mind.

Another correlate of later theory of mind is parents’ and siblings’ use of mental state
terms. For instance, parents’ use of mental state terms when children are 2 years of age
correlates with the children’s usage a year later (Furrow, Moore, Davidge, & Chiasson, 1992), and the frequency of children's mental state talk in conversations with older siblings and mothers at 33 months correlates with false belief at 40 months (Dunn, Brown, Slomkowski, Tesla, & Youngblade, 1991). That this exposure plays a causal role in facilitating theory-of-mind understanding is supported by Ruffman et al. (2002), who showed that mothers’ use of mental state talk consistently predicted later theory-of-mind understanding across three time points over a year, even after controlling for children's own mental state language, their earlier theory-of-mind understanding, their language ability, their age, mother’s education and other types of mother talk. However, theory of mind never predicted mother mental state talk at any subsequent time-point (i.e. it was not a reciprocal relation).

Thus, both a child's developing language ability and their exposure to mental state talk facilitates later theory of mind. That these are in some sense distinct is also shown by Ruffman et al. (2002). They examined contributions of earlier language ability, as measured by the Linguistic concepts subtest of the Clinical Evaluation of Language Fundamentals–Preschool ( CELF–Preschool; Wiig, Secord, & Semel, 1992) to later theory of mind after controlling for earlier mothers' mental state talk (and earlier theory of mind, etc.). They also examined the contribution of earlier mother mental state talk to later theory of mind after controlling for earlier language ability (and earlier theory of mind, etc.). Mother’s mental state talk correlated across all three sets of time points with an average partial correlation of $pr = .36$. Language ability correlated across two sets of time points with an average partial correlation of $pr = .29$.

Although the finding for the role of language ability by Ruffman et al. (2002) is consistent with the claim by Astington and Jenkins (1999) that early child language correlates with later false belief but not the reverse, there are two reasons to be cautious. First, there is evidence that likely precursors of a theory of mind play a role in language understanding. For instance, children’s understanding of intention likely assists them in understanding which object an adult is referring to, hence providing a putative link between mental state understanding and vocabulary development (Baldwin & Moses, 2001). Second, as Astington and Jenkins note themselves, it is possible that a ‘third’ or ‘underlying’ variable such as working memory can account for the perceived relation of language ability and false belief. If language and false belief are both manifestations of the same underlying structure (e.g. working memory), then it is possible that the relation of language to false belief does not reflect a causal relation. Rather, it may simply reflect the fact that language measures are a better measure of this structure than false belief. In favour of this idea, a number of studies have revealed correlations between working memory ability and false belief performance, using a wide variety of tasks (Davis & Pratt, 1995; Gordon & Olson, 1998; Hughes, 1998a, 1998b; Keenan, 1998; Keenan, Olson, & Marini, 1998), and also between language development and working memory (e.g. Adams & Gathercole, 1996; Baddeley, Gathercole, & Papagno, 1998).

Thus, our first aim was to examine Astington and Jenkins’ (1999) claim for an exclusive link between early language and later false belief understanding by testing preschool children at two time points. The second aim was to examine whether working memory could account for this relation by including a modified version of the Backwards Digit Span task used by Davis and Pratt (1995).

The third aim of this study concerns what aspects of language ability are involved. A number of researchers have claimed that it is primarily syntax that plays the crucial role. These claims rest on the perceived similarity of structure underlying syntactic
understanding and false belief reasoning. That is, unlike semantics, which involves the understanding of words, syntax requires following and understanding how the arrangement or combination of words affects the meaning of the sentence. For example, whereas most 3-year-olds will know the individual words of the sentence, ‘The car hit the truck’, correctly representing the sentence requires understanding the importance of word order: ‘car → hit → truck’ not ‘truck → hit → car’. Astington and Jenkins (1999) argue this syntactic ability of keeping track of, and correctly representing, often quite complex relations between individual elements of a sentence is just the type of ability that would help a child keep track of and represent the complex relation between the location of the object and presence or absence of the character in a false belief task. In support of this, they show that items from the TELD said to tap syntax, uniquely predict later false belief over and above semantic items, but the reverse does not hold.

Others have focused on the relation between being able to handle more complex syntax, such as the embedding of a clause into a main clause or sentence. Smith et al. (2003) claim that children’s ability to follow relative clauses, such as, ‘The girl kicked the man that jumped over the wall’, predicts false belief understanding whereas similar, but unembedded clause sentences do not. In contrast, de Villiers and her colleagues argue that it is only a specific type of embedded syntax that plays the crucial role in meta-representational ability, sometimes called object or sentential complements (deVillers & deVillers, 2000; Hale & Tager-Flusberg, 2003), but more recently called tensed object complements (deVillers & Pyers, 2002).

Although syntax and semantics will always be confounded to some extent, it is possible to produce sentences that clearly differ in their relative syntactic and semantic demands. For example, Caramazza and Zurif (1976) showed that aphasic patients were much worse on sentences like, ‘The zebra is chasing the horse’ (where semantic knowledge of horses and zebras does not help in deciding which is chasing which), than on sentences like, ‘The cat is chasing the mouse’ (where semantic knowledge of cats and mice does help in deciding which is likely to be chasing the other). Thus, ‘The zebra is chasing the horse’ can be said to have relatively high syntactic and low semantic demands, and to be primarily a test of syntax. Conversely, vocabulary measures such as the BPVS (British Picture Vocabulary Scale; Dunn, Dunn, Whetton, & Pintillie, 1982), where the child is asked to ‘point to arrow’ (selecting from four pictures), can be said to have relatively low syntactic and high semantic demands, and to be primarily a test of semantics.

Although Astington and Jenkins (1999) found that the ‘syntax’ scale of the TELD was more highly related to false belief than the semantics scale, only one of the items in the TELD syntax scale (‘The truck hit the car’) distinguishes syntax and semantics clearly (see Ruffman et al., 2003). For example, four syntax items require the child to remember and repeat back a sentence, yet this is exactly what the Stanford-Binet Verbal Memory Test requires of children, and this is considered a measure of general language ability (Jenkins & Astington, 1996). Indeed, later versions of the TELD do not make a syntax/semantics distinction between items.

We isolated syntactic and semantic abilities (as far as is possible) using two tests tapping mainly semantics, and two tests tapping mainly syntax. The Semantics: BPVS Test tapped vocabulary (i.e. understanding of non-mental state nouns and verbs). For example, children were asked to ‘point to forest’. Recall that previous research has found correlations between the BPVS and false belief. The Semantics: Linguistic Concepts Test (taken from the Linguistic concepts subtest of the CELF–Preschool) tapped understanding of sentences rather than single words. For example, children
were asked to 'point to the first elephant in line'. Because false belief requires children to understand words in the context of longer utterances rather than words in isolation (as in the BPVS), it was conceivable that the Linguistic Concepts Test would be more highly correlated with false belief than the BPVS.

The Syntax: Word Order Test (adapted from Caramazza & Zurif, 1976) tapped syntactic understanding of word order. For example, children were asked to 'point to the girl jumped over the boy'. The Syntax: Embedded Clause Test tapped children's understanding of centre-embedded clauses. For example, children were asked to 'point to the pencil that is under the box is blue'. The semantic demands were low in that children would know the meaning of each word (pencil, under, box, blue) but could only get the item right by understanding the embedding of the clause and relating it to the main clause (i.e. the pencil is blue not the box).

Thus, we had four language tests each aimed at testing specific aspects of syntax and semantics. These tests were based on a study by Ruffman et al. (2003) who found that semantics was as likely to correlate with false belief as syntax. However Ruffman et al.'s study is inconclusive. In Experiment 1 their design was longitudinal but they only examined early language and later false belief, but not early false belief and later language. In Experiment 2 their design was cross-sectional so that it was not possible to evaluate causal direction. In the present study we use a fully cross-lagged design to compare the unique contribution of syntax and semantics by testing whether particular tests at Time 1 contributed uniquely to the variance in false belief performance at Time 2, after accounting for the contribution of other tests. Recall that that is what Astington and Jenkins found, but their tests of syntax and semantics were very much confounded. If syntactic ability is important we would expect the syntactic tests to predict false belief even after the shared variance with semantic ability has been accounted for.

Because we had separate measure of syntax and semantics, our fourth aim was to consider whether earlier theory of mind made differential contributions to these measures (at the later time-point). Although it is known that understanding of intention is linked to children’s learning the meaning of words (e.g. Baldwin & Moses, 2001), no-one, to our knowledge, has looked at theory of mind facilitating syntactic development, or compared this relation to theory of mind and later semantic understanding.

The fifth and final aim of our study was to examine the causal relation between false belief and working memory. Although research has clearly established a relation between working memory and false belief at a single time point (see above), only one study has employed a longitudinal design. Hughes, (1998a, 1998b) gave a working memory task as part of a battery of ‘executive function’ tasks, together with false belief tasks and a language measure to children with a mean age of 3 years, 11 months, and again 1 year later. Working memory correlated with later false belief, and false belief correlated with later working memory (correlations roughly similar). However, because of the relatively advanced age of her sample, Hughes used advanced measures of false belief at the second time-point. Also, although she showed that false belief did not predict later working memory (partiailing out earlier working memory), no similar analysis was reported for working memory as a predictor of later false belief. Working memory was considered only as a part of a composite of executive function tasks. What was reported was that after partialling out earlier age, language, false belief, and other executive tasks, working memory did not predict later false belief. Our longitudinal data permits a re-examination of whether working memory has a unique causal effect on
false belief over time, by focusing solely on working memory and using a younger sample that had the same false belief tasks at each time-point.

In summary, we aimed to investigate claims regarding the role of language in false belief understanding by comparing: (a) general language ability (syntax + semantics) at Time 1 as a predictor of false belief at Time 2 (after accounting for earlier age and false belief) to (b) false belief at Time 1 as a predictor of general language at Time 2 (after accounting for earlier age and language). We aimed to extend this finding by considering whether working memory mediated this relation. An additional aim was to investigate whether tests of syntactic ability uniquely contribute to false belief by comparing language items with high syntax and low semantic demands, to items with high semantic and low syntax demands. We also aimed to examine the reverse relation, that is, whether false belief differentially facilitated syntactic and semantic development. Lastly, we looked at the relations between working memory and false belief.

Method
Participants
Forty-four preschool children (25 boys, 19 girls) took part in this study. Their mean age at the first time of testing was 3.8 years (range 3–4.3 years). All children spoke English as a first language and were drawn from four nurseries in a predominately middle-class area of a small city in the UK.

Materials and procedure
Children were tested at the end of the autumn term and again during the summer term, approximately 6 months later. Children were tested individually in a quiet area of the main nursery room. Testing took about 20 minutes, including the warm-up session.

Warm up
The warm-up involved pointing at five sheets with pictures on them. The first sheet had pictures of a shoe, sock, bowl, ball, bucket, box, square, and triangle. Children were asked to ‘point to the sock’, and so forth. The second sheet was split into two halves, each with two pictures. The first half showed two tables: one with a ball under a table, and one with a ball above the table. The second half again showed two tables, one with a square above a table, and one below a table. Children were shown one half at a time and asked to ‘point to the square/ball above/under the table’. The third sheet had a blue square and a red square. Children were asked, ‘What colour is this?’, while the experimenter pointed to the red/blue square. The fourth and fifth sheets had pictures of animals on them, taken from the training trials sheets of the Linguistic concepts subtest. The pictures were of a monkey, cat, tiger, elephant, turtle, fish, dog, bear, giraffe, and bird. All children were correct on all questions. Thus all children participating demonstrated that they understood the instruction to point to pictures and knew the key words (e.g. square, red, bear, under, etc.) used in the language tests.

Language tests
There were four language tests given (see Appendix A for a complete list). Each aimed at assessing an aspect of either syntactic or semantic ability. Each test consisted of 10
separate sentences, making 40 language items in total. Each item included four pictures on a sheet of paper, each picture depicting a different interpretation of the test sentence, except for the Semantics: Linguistic concepts subset which used one picture per item. After the warm-up, the child was told ‘I am going to show you some pictures and I want you to point at what I say’. If the first item involved selecting from four pictures (i.e. all the tests except Semantics: Linguistic concepts), the experimenter then said: ‘OK, only one of these pictures is the same as what I say’. The experimenter then pointed to the four pictures for the first test item to ensure the child scanned each and said: ‘Point to what I tell you, point to . . .’, reading out the first target sentence. If the first item was from Semantics: Linguistic concepts the child was told: ‘Look at this picture, now I want you to just point to what I say’. The experimenter then pointed out the elements of the picture and said: ‘Point to . . .’, giving the target sentence’. For each new language item, the experimenter again pointed out the four pictures or individual elements and said: ‘Point to . . .’, giving the target sentence or word. The position of the correct picture (i.e. top left, top right, bottom left, bottom right of the page) was equally distributed within each test (except Semantics: Linguistic concepts) to avoid a pointing bias having differential effect in a particular test.

The Semantics (BPVS) test was taken from the long form of the BPVS. This tests children’s vocabulary, that is, their semantic understanding of basic nouns and verbs. For example, children were asked to ‘point to delivering’ (selecting from pictures of a post-woman handing a package to a child, two children walking up to a door, a mother pushing a child on a swing, and children getting books down from a shelf). The 10 items chosen were selected as being most appropriate for 4-year-olds (items 23 to 32 of the long form; alphas at Times 1 and 2 = .41, .64).

The Semantics: Linguistic Concepts Test was taken from the CELF–Preschool (Wiig et al., 1992). This tests children’s semantic understanding of sentences. For example, children were asked to ‘point to some of the tigers’ (selecting from a picture showing three tigers together with two bears beside them). The 10 items chosen were selected for being lowest in memory demands (αs = .69, .59).

The Syntax: Word Order Test was derived from items used by Caramazza and Zurif (1976). This tests children’s syntactic understanding of word order. For example, children were asked to ‘point to the horse is chasing the zebra’ (selecting from pictures showing a horse chasing a zebra, a zebra chasing a horse, a zebra and horse standing next to each other, and a horse and zebra facing each other; αs = .66, .60). For this test, it was necessary to understand the order of the words (syntax) ‘horse’ and ‘zebra’ in the sentence, in addition to understanding the meaning of the individual words themselves (as in the semantics tests).

The Syntax: Embedded Clause Test tested children’s understanding of centre-embedded clauses. For example, children were asked to ‘point to the square that is under the shoe is red’ (selecting from pictures of a red square under a white shoe, a red shoe above a white square, a white square above a red shoe, and a white shoe under a red square; αs = .42, .50). For this test, it was necessary to understand how word order (syntax) necessitates that it is the square that is red, not the shoe.

Both Caramazza and Zurif (1976) and Ruffman et al. (2003) found that the items of the Syntax: Word Order Test (e.g. ‘The zebra is chasing the horse’) were significantly harder than control items (e.g. ‘The bird is eating the bug’) in which semantic knowledge of birds and bugs would make it less essential to take into account word order. Likewise, the Syntax: Embedded Clauses Test was based on Ruffman et al. who found that such items (e.g. ‘The ball that is under the table is blue’) were significantly
harder than semantically identical but syntactically easier items (e.g. The blue ball is under the table). Thus, both syntax tests required children to understand the importance of word order to sentence comprehension. In contrast, the Semantics: BPVS Test required children to understand the meaning of words in isolation (where word order was not applicable), and the Semantics: Linguistic Concepts Test required them to understand the meaning of words in a sentence but where word order (i.e. the order of two nouns) was not crucial to understanding. For this reason, the syntax tests unequivocally placed higher demands on understanding word order and syntax.

**Theory of mind**

Three false belief tasks were used at both time points: two unexpected transfer tasks (Wimmer & Perner, 1983) and one unexpected contents task (Perner, Leekam, & Wimmer, 1987). The unexpected-transfer tasks all involved a story character placing an object in one location, leaving, and another character moving the object to another location while the first character was away. When the first character returned, the child was asked the belief question (‘Where will [story character] look for her [object] first?’), a justification question (‘Why will [story character] look there?’), and two control questions to ensure the child followed the story (‘Where was the [object] in the beginning?’ and ‘Where is the [object] now?’). The unexpected-contents task involved asking the child what was in a familiar box. After they answered correctly, the box was opened and shown to have unexpected contents inside. The box was closed again and the child asked again what was inside. Then they were asked the ‘self’ belief question (‘When I first showed you this box, all closed up like this, what did you first think was inside?’) and an ‘other’ belief question about a doll who had not seen inside (‘When we first show [story character] this box, before she looks inside, what will she think is inside?’). They were also asked a justification question: ‘Why will she think there are [what child said] in there?’).

Playmobile® figures were used for the characters. At Time 1 a doll’s-house size cupboard and cake were used in one false belief transfer task, and two coloured cups (pink and red) with lids, and a ball, were used for the other transfer task. A different cupboard and different containers (green and orange boxes) were used at Time 2. A crayon (Crayola®) box containing a key was used for the unexpected contents task at Time 1 and a Smarties® box containing a pencil was used at Time 2.

**Working memory**

The working memory task was a modified Backwards Digit Span task, adapted from Davis and Pratt (1995). Instead of numbers (which have variable familiarity to such young children), children were asked to repeat backwards sequences of familiar words (taken from Snodgrass & Vanderwart, 1980 and training items of the BPVS). The children received a training phase where pictures were placed in a line in front of the child as they were named. The experimenter pointed to the pictures in a backwards sequence and named the pictures to help children understand how to say the sequence backwards. Thus the experimenter said, ‘I say “teddy, ball” as pictures of a teddy and ball were laid out in front of the child, and then, ‘You say those backwards, you say . . . ’ leaving the child to name the pictures in a backwards sequence, prompted by the experimenter pointing at the pictures in the backwards order. Once the child was familiar with saying the sequence backwards, the experimenter said, ‘Now I am not
going to put pictures down, I am just going to say the words. You say what I say in a backwards way'. The child then received the test sequences: three with two words, and three with three words. If the child was incorrect on the first two test trials, they were again shown how to say the words backwards with pictures (see Appendix B for full details).

Design
The 40 language items of the four tests were randomized and split into three roughly equal sets. These sets were counterbalanced into three fixed orders which meant that each set appeared at the beginning, middle, and end of the testing session in roughly equal amounts. The false belief tasks were given in a block and counterbalanced with the working memory task. Thus the order of presentation was language set – working memory or false belief – language set – false belief or working memory – language set.

Scoring
Language
Raw scores were used for the language tests, with a maximum score of 10 for each test. A composite score for general language ability was created by summing the raw scores for the four tests, making a maximum possible score of 40. The internal consistency (α) of the composite score at Time 1 was .76 and at Time 2 was .79.

Theory of mind
Each transfer task yielded two points (one for the belief question, one for the justification of the belief question). The contents task yielded three points (one for the ‘self’ belief question, one for ‘other’ belief and one for the justification of the other question). For the transfer task, justifications were scored as correct if they referred to the original location of the object, or to the story character thinking the object had not moved or had not seen it moved. For the contents tasks, correct justifications referred to the box’s misleading appearance (Clements, Rustin, & McCallum, 2000; Wimmer & Mayringer, 1998). Control questions had to be answered correctly for credit to be given for the test questions. Correct justifications were only counted if the child correctly answered the belief (and control) questions. The internal consistency (α) of the composite theory-of-mind measure was .89 at Time 1 and .83 at Time 2.

Working memory
One point was given for correctly reversing two words, two points for correctly reversing three words. Half points were given for reversing two words that were not adjacent (although these were infrequent). The maximum score possible was nine (three points for reversing three sets of two words, six points for reversing three sets of three words).

Results
Three children at Time 1 passed the test question in a transfer task but failed a control question. These children were scored as failing the test question. No child failed
a control question at Time 2. All children correctly remembered the unexpected contents of the box before being asked test questions at both time points. Table 1 gives the descriptive data for the language, theory of mind and working memory tasks at both time points.

### Table 1. Means, standard deviations and ranges of age and tasks at both time points

<table>
<thead>
<tr>
<th></th>
<th>Time 1</th>
<th>Time 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Age</td>
<td>3.80</td>
<td>0.33</td>
</tr>
<tr>
<td>Theory of mind</td>
<td>3.18</td>
<td>2.58</td>
</tr>
<tr>
<td>Language (4 tests composite)</td>
<td>26.00</td>
<td>5.35</td>
</tr>
<tr>
<td>Semantics: BPVS</td>
<td>6.59</td>
<td>1.78</td>
</tr>
<tr>
<td>Semantics: linguistic concepts</td>
<td>6.75</td>
<td>2.07</td>
</tr>
<tr>
<td>Syntax: word order</td>
<td>8.25</td>
<td>1.86</td>
</tr>
<tr>
<td>Syntax: embedded clause</td>
<td>4.41</td>
<td>1.92</td>
</tr>
<tr>
<td>Working memory</td>
<td>0.76</td>
<td>1.22</td>
</tr>
</tbody>
</table>


All individual language measures, false belief composite and working memory measures increased significantly between Time 1 and Time 2 (all paired-sample t tests, \( p < .01 \), one-tailed, which are still significant with Holms correction for six \( t \) tests). Mean scores for individual language measures were significantly above chance (all one-sample \( t \) tests, \( p < .001 \)). Means and standard deviations for language tests (individual and composite) and false belief performance show that there were no floor or ceiling effects. Children found the working memory task difficult but their performance was in line with previous research (Davis & Pratt, 1995; Gordon & Olson, 1998). Re-scoring as a dichotomous variable (i.e. scoring 0 if the child never reversed words and one if they did reverse at least two words one or more times) shows that 17 children (39%) at Time 1 ‘passed’ working memory. This proportion is similar to Davis and Pratt (1995) and Gordon and Olson (1998). Because reducing working memory to a dichotomous variable reduces sensitivity, and our aim was to maximize sensitivity and variance, we used the continuous scale (although using dichotomous scores yielded the same pattern of results). There were no order or gender effects for any of the tasks at either time point.

The raw correlations (and important partial correlations) between general language ability (a composite of scores from each of the four individual language tests), theory of mind, working memory and age are shown in Table 2. Within each time-point, language and theory of mind correlated well, replicating earlier work (see above). Working memory and theory of mind also correlated well, again replicating previous work. Between time points, language and theory of mind again correlated well, replicating Astington and Jenkins (1999) and Ruffman et al. (2002). Like Hughes (1998a, 1998b) we found a good correlation between earlier false belief and later working memory but (unlike Hughes) the reverse relation failed to reach significance (though it was in the expected direction). Language at Time 1 and language at Time 2 correlate highly, as do theory of mind at Time 1 and theory of mind at Time 2, again replicating Astington and
Jenkins, and supporting claims for stability of these abilities. Likewise, working memory at Time 1 correlated very well with working memory at Time 2, replicating Hughes, and again supporting the idea of a stable construct.

To address the first aim of this investigation, namely the relation between general language and theory of mind, two hierarchical regressions were carried out. Table 3 reports data from the final step of each regression. The first regression, shown in the top

Table 2. Correlations (and partial correlations) within and between time points

<table>
<thead>
<tr>
<th></th>
<th>ToM1</th>
<th>ToM2</th>
<th>Lang1</th>
<th>Lang2</th>
<th>WM1</th>
</tr>
</thead>
<tbody>
<tr>
<td>ToM1</td>
<td>.42**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ToM2</td>
<td>.25* (.03)</td>
<td>.68*** (.51***)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang1</td>
<td>.36**</td>
<td>.60***</td>
<td>.63*** (.44***)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang2</td>
<td>.22 (.11)</td>
<td>.58*** (.21)</td>
<td>.54***</td>
<td>.79*** (.66***)</td>
<td></td>
</tr>
<tr>
<td>WM1</td>
<td>.35**</td>
<td>.37***</td>
<td>.20 (.22)</td>
<td>.46**</td>
<td>.30* (.16)</td>
</tr>
<tr>
<td>WM2</td>
<td>.24</td>
<td>.41***</td>
<td>.34*</td>
<td>.53***</td>
<td>.47***</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001, all one-tailed.

Note. ToM1 = Theory of mind at Time 1, ToM2 = Theory of mind at Time 2, Lang1 = Language (composite) at Time 1, Lang2 = Language (composite) at Time 2, WM1 = Working memory at Time 1, WM2 = Working memory at Time 2. Partial correlations are in brackets. Partial correlations with ToM2 are independent of age, ToM1, WM1 and Lang1 (e.g. the partial correlation between ToM1 and ToM2 is shown having partialled out age, WM1 and Lang1). Partial correlations with Lang2 are independent of age, Lang1, WM1 and ToM1 (e.g. the partial correlation between WM1 and Lang2 is shown having partialled out age, Lang1 and ToM1).

Jenkins, and supporting claims for stability of these abilities. Likewise, working memory at Time 1 correlated very well with working memory at Time 2, replicating Hughes, and again supporting the idea of a stable construct.

To address the first aim of this investigation, namely the relation between general language and theory of mind, two hierarchical regressions were carried out. Table 3 reports data from the final step of each regression. The first regression, shown in the top

Table 3. Hierarchical regression analyses for language (at Time 1) predicting theory of mind (at Time 2) and for theory of mind (at Time 1) predicting language (at Time 2)

<table>
<thead>
<tr>
<th>Variable</th>
<th>β</th>
<th>t</th>
<th>R²/ΔR²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age1</td>
<td>−.03</td>
<td>−.21</td>
<td></td>
</tr>
<tr>
<td>ToM1</td>
<td>0.50</td>
<td>3.67***</td>
<td>.46***</td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WM1</td>
<td>−.17</td>
<td>−1.43</td>
<td>.00</td>
</tr>
<tr>
<td>Step 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lang1</td>
<td>0.42</td>
<td>3.02***</td>
<td>.10*</td>
</tr>
<tr>
<td>Time 1 ToM to Time 2 language Step 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age1</td>
<td>0.08</td>
<td>0.71</td>
<td></td>
</tr>
<tr>
<td>Lang1</td>
<td>0.70</td>
<td>5.51***</td>
<td>.61***</td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WM1</td>
<td>−.11</td>
<td>−1.00</td>
<td>.00</td>
</tr>
<tr>
<td>Step 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ToM1</td>
<td>0.07</td>
<td>1.32</td>
<td>.02</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001. All figures taken from the final step.

Note. β = Standardized regression weight, R²/ΔR² = Amount of variance/change in amount of variance accounted for, Age1 = Age at Time 1, ToM1 = Theory of mind at Time 1, Lang1 = Language (composite) at Time 1, WM1 = Working memory at Time 1.
half of the table, clearly shows that language significantly predicts later false belief performance, even after accounting for earlier age, false belief ability, and working memory (accounting for an additional 10% of later false belief variance). The second regression, however, shown in the bottom half of the table, shows the reciprocal relation does not hold: false belief performance does not significantly predict later language ability, after accounting for earlier age, language understanding, and working memory (accounting for only an additional 2% of later language variance). In this respect we replicate the finding by Astington and Jenkins (1999) that language ability significantly predicts later theory of mind (but not vice versa), and extend it (the second aim) by showing that working memory, at least as measured by the modified Backwards Digit Span task, does not mediate the language - theory of mind relation.

However, in another respect we do not replicate Astington and Jenkins' basic finding. This concerns the issue of sensitivity of language and theory-of-mind measures. Language measures typically have many items and so produce a wide range of scores. Theory-of-mind measures, even when combined to form a composite measure, typically have many fewer items. Thus theory-of-mind measures may be less sensitive than language measures and lower sensitivity may mask any true effects that theory of mind has on later language ability.

One way to equate language and theory of mind sensitivity is to take as many items from the language composite as there are points available from the theory of mind composite, and to use this score as the language measure, rather than the full 40-item measure. We could equate the language sample and false belief sample in two ways. The first would be to take the full composite score of false belief that includes justifications. Justifications have been used in some studies as a stronger measure of false belief understanding (de Villiers & Pyers, 2002, p. 1048). Including them also increases the range and hence sensitivity of scores and gives a possible score (7) that is comparable with the possible score of 6 used by Astington and Jenkins (1999). The other way to equate language and false belief would be to use the false belief composite without scores for correct justifications (i.e. using a possible score out of 4 instead of 7). This reduces the range of scores possible (and hence sensitivity), but it does make the false belief composite comparable to that of Astington and Jenkins because they did not use justifications. Further, it may be that including the more linguistically demanding justification question obscures the true relation between false belief and language.

We decided to equate language and false belief scores both ways. That is we took samples of seven language items and compared them to the full false belief composite (with justifications) and we also took samples of four language items and compared them to the false belief composite without justifications. Moreover, we sampled each of these ways a large number of times to avoid sampling biases.

Thus, we first took 100 random samples of seven language items from the 40 available at Time 1, and entered them into 100 regressions predicting later false belief with justifications (after partialling out age, earlier false belief, and working memory). The mean $R^2$ for these regressions was .46. The mean $\Delta R^2$ for language in these regressions was .049 ($SD = 0.04$) with a 90% confidence interval between .005 and .126. Of these 100 regressions, 42 were significant (that is, in 42 out of the 100 regressions, language significantly predicted later false belief over and above the contribution of age, earlier false belief, and working memory). Language explained an average of 5% of the unique variance over and above that accounted for by earlier age, false belief, and working memory. A similar result was found, however, running 100 regressions the other way, with earlier false belief predicting later random samples of
seven language items (after partialling out earlier age, the earlier seven language items, and working memory). In this case, the mean $R^2$ for these regressions was .31 (with a 90% confidence interval between .164 and .489). The mean $\Delta R^2$ for false belief in these regressions was .051 ($SD = 0.03$) with a 90% confidence interval between .003 and .110. Of the 100 regressions, 25 were significant. Thus, false belief significantly predicted later language, contributing an average of 5% to later variance over and above that accounted for by earlier age, language (seven items), and working memory.

We next considered the false belief composite without justifications. We took 100 random samples of four language items from the 40 available at Time 1, and entered them into 100 regressions predicting later false belief without justifications (after partialling out age, earlier false belief, and working memory). The mean $R^2$ for these regressions was .374. The mean $\Delta R^2$ for language in these regressions was .024 ($SD = 0.029$) with a 90% confidence interval between .000 and .086. Of these 100 regressions, 12 were significant, with language explaining an average of 2.4% of the unique variance over and above that accounted for by earlier age, false belief, and working memory. A stronger result was found, however, running 100 regressions the other way, with earlier false belief predicting later random samples of four language items (after partialling out earlier age, the earlier four language items, and working memory). In this case, the mean $R^2$ for these regressions was .179 (with a 90% confidence interval between .057 and .338). The mean $\Delta R^2$ for false belief predicting later language (four items) in these regressions was .07 ($SD = 0.05$) with a 90% confidence interval between .007 and .168. Of the 100 regressions, 36 were significant. Thus, false belief (without justifications) significantly predicted later language, contributing an average of 7% to later variance over and above that accounted for by earlier age, language (four items), and working memory.

What is clear, then, is that sampling affects the relation between language and theory of mind. Sampling seven items and comparing to a false belief composite with justifications shows that language is a more reliable predictor of later false belief than false belief is of later language. Sampling four items, however, and comparing to a false belief composite without justifications shows that false belief is a more reliable predictor of later language than language is of later false belief. Although sampling clearly makes a difference, the consistent finding in both cases, however, was that false belief predicted later language, indicating that the relation between language and false belief is bidirectional.

We now turn to individual language tasks tapping syntax and semantics. Table 4 lists the raw correlations (and important partial correlations) between each individual language test and false belief (and correlations with each other). There was a general trend for language and theory of mind tests to correlate with one another over time points. Interestingly, the semantic tests (5/6 correlations significant = 83%) and syntactic tests (5/6 correlations significant = 83%) were not more likely to correlate with one another, in comparison to correlations between the semantic and syntactic tests (13/16 correlations significant = 81%). This is despite the fact that the syntax tests unequivocally place higher demands on word order and syntactic ability than the semantic tests. This result replicates Ruffman et al. (2003).

The third aim of this study was to consider the unique contribution of each separate language test to later false belief performance. For this purpose, four separate hierarchical regressions were conducted (see Table 5). Earlier age and false belief were entered in the first block, then all other language tests were entered (i.e. three out of four) in the next block, and finally the language test under consideration was entered. Having accounted for the other three language tasks, no language task was a significant
Table 4. Correlations (and partial correlations) of age, theory of mind, working memory and language subtests within and between time points

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>ToM1</th>
<th>ToM2</th>
<th>Semant: BPVS1</th>
<th>Semant: BPVS2</th>
<th>Semant: LC1</th>
<th>Semant: LC2</th>
<th>Syntax: Word Ord.1</th>
<th>Syntax: Word Ord.2</th>
<th>Syntax: Embed1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Semant: BPVS1</td>
<td>.27*</td>
<td>.39**</td>
<td>.48**</td>
<td>(.29)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semant: BPVS2</td>
<td>.10</td>
<td>.46**</td>
<td>.44**</td>
<td></td>
<td>.41**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semant: LC1</td>
<td>.09</td>
<td>.47**</td>
<td>.53**</td>
<td>(.31)</td>
<td></td>
<td>.22</td>
<td>.33*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Semant: LC2</td>
<td>.37**</td>
<td>.58**</td>
<td>.52**</td>
<td>(.28)</td>
<td>.32*</td>
<td>.49**</td>
<td>.56***</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syntax: Word Ord.1</td>
<td>.24</td>
<td>.53**</td>
<td>.43**</td>
<td>(.06)</td>
<td>.36**</td>
<td>.59***</td>
<td>.45**</td>
<td>.54***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syntax: Word Ord.2</td>
<td>.00</td>
<td>.19</td>
<td>(.20) .25</td>
<td>.06</td>
<td>.28*</td>
<td>.15</td>
<td>.33*</td>
<td>.61***</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Syntax: Embed1</td>
<td>.42**</td>
<td>.29**</td>
<td>.34*</td>
<td>(.11)</td>
<td>.36**</td>
<td>.49**</td>
<td>.24</td>
<td>.32*</td>
<td>.20</td>
<td></td>
</tr>
<tr>
<td>Syntax: Embed2</td>
<td>.16</td>
<td>.42**</td>
<td>.33**</td>
<td>(.32)</td>
<td>.26*</td>
<td>.36**</td>
<td>.42**</td>
<td>.42**</td>
<td>.50***</td>
<td>.34*</td>
</tr>
</tbody>
</table>

*p < .05, **p < .01, ***p < .001, all one-tailed.

Note. ToM1 = Theory of mind at Time 1, ToM2 = Theory of mind at Time 2, Semant: BPVS1 = Semantics: BPVS subtest at Time 1, Semant: LC1 = Semantics: linguistic concepts subtest at Time 1, Syntax: Word Ord.1 = Syntax: word order subtest at Time 1, Syntax: Embed1 = Syntax: embedded subtest at Time 1, Semant: BPVS2 = Semantics: BPVS subtest at Time 2, Semant: LC2 = Semantics: linguistic concepts subtest at Time 2, Syntax: Word Ord.2 = Syntax: word order subtest at Time 2, Syntax: Embed2 = Syntax: embedded subtest at Time 2. Partial correlations with ToM2 are shown independent of age at Time 1, ToM1, and all other language tests at Time 1 (e.g. the partial correlation between semantics: BPVS1 and ToM2 is shown having partialled out Age1, ToM1, Semantics: linguistic concepts1, Syntax: word order1, and Syntax: embedded1). Partial correlations between ToM1 and each language measure at Time 2 are shown independent of age at Time 1 and that language measure at Time 1 (e.g. the partial correlation between ToM1 and Semantics: BPVS2 is shown independent of Age1 and semantics: BPVS1).
Table 5. Summary of hierarchical regression analysis predicting theory-of-mind scores at Time 2 from each language subtest measure at Time 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>β</th>
<th>t</th>
<th>(R^2/\Delta R^2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>For all regressions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.04</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>ToM</td>
<td>0.49</td>
<td>3.33**</td>
<td>0.46***</td>
</tr>
<tr>
<td>For Semantics: BPVS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 2 All other language tests</td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>Step 3 Semantics: BPVS</td>
<td>0.23</td>
<td>1.85*</td>
<td>0.04</td>
</tr>
<tr>
<td>For Semantics: Linguistic concepts</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 2 All other language tests</td>
<td></td>
<td></td>
<td>0.07</td>
</tr>
<tr>
<td>Step 3 Semantics: Linguistic concepts</td>
<td>0.25</td>
<td>1.96*</td>
<td>0.04</td>
</tr>
<tr>
<td>For Syntax: Word order</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 2 All other language tests</td>
<td></td>
<td></td>
<td>0.11*</td>
</tr>
<tr>
<td>Step 3 Syntax: Word order</td>
<td>-0.05</td>
<td>-0.35</td>
<td>0.00</td>
</tr>
<tr>
<td>For Syntax: Embedded</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 2 All other language tests</td>
<td></td>
<td></td>
<td>0.11*</td>
</tr>
<tr>
<td>Step 3 syntax: Embedded</td>
<td>0.09</td>
<td>0.71</td>
<td>0.00</td>
</tr>
</tbody>
</table>

\*p < .05, **p < .01, ***p < .001, \'p < .07.

Note. ToM1 = Theory of mind at Time 1.
All figures are taken from the final step.

Discussion

This study investigated two recent and related claims for the role of language in theory-of-mind development in typically developing preschool children. The first claim, that predictor of false belief performance on its own, although the two semantics tasks were both marginally significant. There was no evidence of either syntax test having a special role in later false belief, over and above the contribution of semantics. Using the false belief composite without justifications produced a near identical pattern of results except that now the semantics tests were not even marginally significant.

We also considered the reverse of this last analysis, that is, the contribution of earlier false belief to each separate language test at Time 2 (see Table 6). Having accounted for age and individual language task at Time 1, early false belief performance was a significant predictor of performance on one of two semantic tests (and a marginal predictor of the other), and one of two syntax tests. To determine whether early theory of mind facilitates individual language skills independent of other language skills, we then carried out identical analyses except that Time 1 performance on all four language tasks was accounted for when examining Time 2 performance on a language task. In these analyses, theory of mind was never a significant predictor of Time 2 performance on an individual language task, never accounting for more than 2% of the variance.

The fifth aim was to consider the relation of working memory and theory of mind. Table 7 shows that working memory does not predict later false belief after accounting for earlier age and false belief. Nor does false belief predict later working memory after accounting for earlier age and working memory. The pattern is the same if earlier language is partialled out as well.
Table 6. Summary of hierarchical regression analysis predicting each language subtest measure at Time 2 from theory-of-mind scores at Time 1

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$R^2/\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>For ToM Time 1 to Semantics: BPVS at Time 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1 Age</td>
<td>-0.04</td>
<td>-0.28</td>
<td></td>
</tr>
<tr>
<td>Semantics: BPVS Time 1</td>
<td>0.28</td>
<td>1.91</td>
<td>.18*</td>
</tr>
<tr>
<td>Step 2 ToM 1</td>
<td>0.37</td>
<td>2.37*</td>
<td>.10*</td>
</tr>
<tr>
<td>For ToM Time 1 to Semantics: LC at Time 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1 Age</td>
<td>0.31</td>
<td>2.58*</td>
<td></td>
</tr>
<tr>
<td>Semantics: LC Time 1</td>
<td>0.41</td>
<td>3.27*</td>
<td>.48***</td>
</tr>
<tr>
<td>Step 2 ToM 1</td>
<td>0.25</td>
<td>1.86*</td>
<td>.04</td>
</tr>
<tr>
<td>For ToM Time 1 to Syntax: word order at Time 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1 Age</td>
<td>0.04</td>
<td>-0.32</td>
<td></td>
</tr>
<tr>
<td>Syntax: word order Time 1</td>
<td>0.67</td>
<td>4.81***</td>
<td>.37***</td>
</tr>
<tr>
<td>Step 2 ToM 1</td>
<td>-0.19</td>
<td>1.25</td>
<td>.02</td>
</tr>
<tr>
<td>For ToM Time 1 to Syntax: embedded clauses at Time 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1 Age</td>
<td>-0.01</td>
<td>-0.09</td>
<td></td>
</tr>
<tr>
<td>Syntax: embedded clauses Time 1</td>
<td>0.37</td>
<td>2.52*</td>
<td>.22**</td>
</tr>
<tr>
<td>Step 2 ToM 1</td>
<td>0.32</td>
<td>2.15*</td>
<td>.08*</td>
</tr>
</tbody>
</table>

* $p < .05$, ** $p < .01$, *** $p < .001$, 'p < .07.
Note. ToM1 = Theory of mind at Time 1.
All figures are taken from the final step.

Table 7. Hierarchical regression analyses for working memory (at Time 1) predicting theory of mind (at Time 2) and for theory of mind (at Time 1) predicting working memory (at Time 2)

<table>
<thead>
<tr>
<th>Variable</th>
<th>$\beta$</th>
<th>$t$</th>
<th>$R^2/\Delta R^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time 1 Working memory to Time 2 Theory of mind</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.01</td>
<td>0.05</td>
<td></td>
</tr>
<tr>
<td>ToM1</td>
<td>0.70</td>
<td>5.32***</td>
<td>.46***</td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WM1</td>
<td>-0.01</td>
<td>-0.52</td>
<td>.00</td>
</tr>
<tr>
<td>Time 1 Theory of mind to Time 2 Working memory</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Step 1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.03</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>ToM1</td>
<td>0.47</td>
<td>3.32***</td>
<td>.32***</td>
</tr>
<tr>
<td>Step 2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WM1</td>
<td>0.23</td>
<td>1.56</td>
<td>.04</td>
</tr>
</tbody>
</table>

* $p < .05$, ** $p < .01$, *** $p < .001$.
Note. Age1 = Age at Time 1, ToM1 = Theory of mind at Time 1, WM1 = Working memory at Time 1.
All figures taken from the final step.
general language ability facilitates later theory-of-mind understanding (but not vice versa) made by Astington and Jenkins (1999), is replicated here but only in a certain respect. Like Astington and Jenkins, we found that language ability, as typically measured with a large number of items, predicts later theory of mind, but that the reverse relation does not hold. Indeed our pattern of results was remarkably similar, even though a different measure of language ability and a different combination of false belief tasks were used. This picture was not so clear-cut, however, when the language measure was reduced to equate the range of scores to the range possible on the theory-of-mind measure. Once this was done, and a sufficient number of analyses were conducted to extensively compare language and false belief; false belief also had a significant effect on later language. For their sets of time points, Astington and Jenkins carried out a single analysis in which they reduced the number of language items to the false belief total. Although they did not find a relation between early false belief and later language, the potential for sampling differences means that a single analysis will be insufficient to provide a sensitive test of the relation between early false belief and later language. Indeed, recall that in the majority of the 100 analyses we conducted, we found that early false belief was not a significant predictor of later language (and that this was also true when going from early language to later false belief). This is not entirely surprising. Even when examined within a single time point, the correlation between language and false belief can be very small (e.g. Taylor & Carlson, 1997) and even non-significant (e.g. Astington & Jenkins, 1999, Frith, Happé, & Siddons, 1994).

In sum, our data are consistent with studies finding that a child’s general linguistic ability directly contributes to their later theory of mind, and also in finding that under certain conditions language is a more consistent predictor of false belief than vice versa. Nevertheless, our data challenge the claim that theory of mind plays no similar role on later language. It seems that task sensitivity, to some extent, masks the role of theory of mind in language development.

This bidirectional relation between language and theory of mind is clearly consistent with claims that changes in mental state understanding facilitate semantic development (Baldwin & Moses, 2001). For instance, Sabbagh and Baldwin (2001) found that a shift in understanding knowledge states between 3 and 4 years of age facilitates new word learning. When a speaker was hesitant but likely to be knowledgeable, 4-year-olds (but not 3-year-olds) tended to learn a novel word pairing. In contrast, when a speaker was hesitant and unlikely to be knowledgeable, 4-year-olds avoided learning a novel word. (See also Birch & Bloom, 2002, and Bloom, 2001, for more on the possible role of theory of mind in children’s learning of the meaning of words.)

Consistent with this idea, we found that false belief was a significant predictor of one of two semantics tests (and a marginal predictor of the other), and one of two syntax tests (after accounting for age and earlier performance on individual language task). However, in no case was false belief a significant predictor of an individual Time 2 language task, after taking into account performance on all Time 1 language tasks. In other words, the relation between early false belief and later language included but was not specific to semantic development (presumably because semantic and syntactic development are highly related in normal development; Ruffman et al., 2003). Our finding thus makes a novel contribution to the research in this area because, as far as we know, no-one has examined such relations before.

Importantly, we also extended the findings of Astington and Jenkins (1999) by showing that working memory, measured using a task similar to that used in previous research, does not account for the relation between language and false belief. This was
possible given the known relation between vocabulary development and working memory and the possibility that syntax is also related to working memory (Gibson, 1998). It is possible, of course, that some other aspect of executive functioning might mediate the relation between language and false belief. For instance, Carlson, Moses, and Breton (2002) have argued that inhibitory efficiency in combination with working memory, rather than working memory alone, is related to false belief understanding. Although this remains possible, there is no clear theory at the moment for how inhibitory ability would be intimately related to language development (and subsequently to false belief).

Concerning the other main aim of this study, that it is syntax that plays the crucial role in the relation between language and false belief, we found no evidence for syntactic understanding (i.e. understanding of word order and embedded clauses) having a unique role in facilitating later theory of mind, over and above that shared by semantics. This longitudinal finding is consistent with, and extends, Ruffman et al. (2003). Likewise, our findings are consistent with Varley and Siegal (2000) who showed that an agrammatic aphasic adult male (S.A.) was able to pass false belief tasks even though he lacked any syntactic ability (e.g. he was unable to understand reversible sentences such as those tapped by our Syntax: Word Order Test). However, because this patient was an adult when he lost syntactic ability it could be argued, as Astington and Jenkins (1999, p. 1312) have, that such a finding does not address whether syntax played a direct role in the development of false belief in the first place. Astington and Jenkins’ argument is clearly challenged by our data.

In all cases, Linguistic concepts was a higher correlate of false belief than the BPVS (see Table 4). This is consistent with the idea that false belief requires children to understand words in the context of longer utterances (rather than single-word utterances as in the BPVS).

de Villiers and Pyers (2002) argue that it is actually a very specific aspect of syntax, tensed object complements, that relates to false belief. Direct tests of children’s understanding of object complements seem to challenge this claim, however. Perner, Sprung, Zauner, and Haider (2003) show that, in German, object complements are understood when used for desires (i.e. want that) much earlier than when used to express belief (i.e. think that) even though the syntactic construction is identical in each case. Likewise, Astington (2000) cites Custer’s (1996) finding that, in English, children understand complements relating to pretence earlier than complements relating to thinking.

In keeping with these findings of more direct tests of de Villiers’ claim, we question whether their data unequivocally supports such a clear case for the unique role of this aspect of syntax. First, the magnitude of the raw correlations between measures of object complements (syntax) and false belief are in line with that which we (and others) have obtained between semantics and belief. Second, unlike us, they provided no comparison between syntax and semantics, so there is no way of directly comparing their results. Third, there is evidence from their own data of a bidirectional relation between syntax and theory of mind. de Villiers and Pyers (2002) report longitudinal data from three time points of 3-year-olds’ understanding of complements (i.e. syntax) and false belief. In support of the claim that the acquisition of this aspect of syntax underpins false belief understanding, they show that being able to answer questions that are said to tap complement understanding predicts false belief performance both at Time 2 and between Time 2 and Time 3, whereas false belief understanding at Time 2 does not predict later complement understanding (at Time 3). However, as they
acknowledge, this is not the case for earlier time points. Between Time 1 and 2, false belief predicts later complement understanding (syntax) equally well as earlier complement understanding predicts later false belief ability.

Fourth, the role of syntax may be inflated in the de Villiers and Pyers (2002) study. This is plausible because, although in their regressions they partial out the child’s productive language ability, they do not partial out earlier false belief understanding. That is, they report longitudinal, but not cross-lagged, relations between false belief and complement understanding. It is essential, we argue, to account for earlier false belief performance, for example, at Time 2, to provide stronger evidence for a causal relation between syntactic ability at Time 2 and false belief at Time 3. This is because the effect of the earlier syntax could be due to shared variance between syntax and false belief at Time 2. In other words, syntax might predict later false belief because (a) false belief correlates with syntax at Time 2, and (b) false belief at Time 2 correlates with false belief at Time 3. Cross-lagging is particularly important in this case, moreover, because of the potential confound between the measure of complement understanding and false belief understanding (Ruffman et al., 2003).

A fifth reason is that complement understanding was tested using sentences such as, ‘She thought she found her ring, but it was really a bottle cap’ and asking, ‘What did she think?’. Researchers have argued that memory for a past statement is mediated by one’s theory of false belief (Perner, 1991; Wimmer & Hartl, 1991). Without such a theory (i.e. that one can think things that are false) the child has no basis for reconstructing what was said, and remembering the mistaken proposition. Instead, the (3-year-old) child’s report of what was said corresponds to their current level of understanding, so they answer in terms of what they know to be true. In sum, the ability to answer correctly (‘it was her ring’) might tap the same understanding (i.e. the understanding of false belief) as required in a false belief task. Moreover, it may be that this understanding is demonstrated earlier in the complementation task because the task places fewer extraneous task demands on the child. For example, the task involves both making the mental state of the protagonist salient and an absence of the actual object in question, two factors that Wellman et al. (2001) show have a significant effect on improving false belief performance.

Taken together, these findings argue for a bidirectional relation between language and theory of mind. For convenience we have referred to our measures of false belief as measures of ‘theory of mind’. We used false belief measures, in part, because arguments that they are unambiguous markers of mental state understanding (Perner, 1991), and also because there are clear theories relating language to false belief. However, we acknowledge that there is more to theory of mind than an understanding of false belief (Bloom & German, 2000) and, indeed, more to language than syntax and semantics (e.g. Siegal, 1999). It is entirely possible that a different pattern of correlations would emerge if other theory of mind measures (e.g. understanding the link between desires and emotions or desires and actions) or other aspects of language (e.g. conversational awareness) were employed.

With respect to syntax and semantics, however, our clear finding was that it is general language – syntax and semantics – that correlates with theory of mind. This is not to say that syntax is unimportant, just that it is no more important than semantic understanding. Indeed, the idea that it is general language ability that is important seems to be the most sensible account of why so many studies using very different language tasks typically all show good correlations with theory of mind. Though these tasks are very different in nature, they provide a good index of general language ability and it is
this general ability, rather than any particular aspect of language, that is important. So, for example, Cutting and Dunn (1999) showed that both receptive vocabulary (as measured by the BPVS) and aspects of expressive narrative speech such as level of grammatical complexity, coherence and length of sentence (as measured by the Bus Story; Renfrew, 1991) all correlate equally well with false belief, even when age is controlled.

Furthermore, understanding propositions (false or otherwise) requires both syntactic and semantic insight. It is important to understand both the difference between ‘pretending X’ and ‘thinking X’ (i.e. mental state semantics), but also between being ‘scared of X’ and being ‘scared of Y’, or ‘thinking X’ rather than ‘thinking Y’ (i.e. non-mental state semantic differences). Tracking these differences arguably draws as much on semantic understanding as syntactic understanding (Olson, 1988; Ruffman et al., 2003). General language (involving syntax and semantics) might assist by allowing children to reflect on and refine a previously existing implicit understanding of false belief. Indeed, there is evidence that language is a more consistent correlate of explicit measures of false belief than measures of eye gaze thought to tap implicit understanding (Ruffman, 2000; Ruffman, Garnham, & Rideout, 2001). This theoretical orientation also seems to be the position recently argued for by Carruthers (2002).

The last aim of our study was to consider the relation between working memory and theory of mind. We replicated Hughes’ (1998a, 1998b) finding that once early working memory is accounted for, false belief does not predict later working memory. Our (consistent) finding was obtained using a different measure of working memory, a younger sample, and identical false belief tasks at each time point (unlike Hughes). We also obtained the novel finding (not examined by Hughes), that working memory does not predict later false belief when early false belief is accounted for. This finding does not contradict cross-sectional studies showing that working memory and false belief improve at the same time (Davis & Pratt, 1995; Gordon & Olson, 1998; Keenan et al., 1998). Indeed, we also obtained within-time-point correlations between false belief and working memory. Such findings are thus consistent with an expression rather than an emergence account of false belief (Carlson et al., 2002). That is, our findings are consistent with the idea that, at a minimum, a certain level of working memory resources might be necessary for success on false belief tasks but raise some doubt over whether improvements in working memory enable children to learn about false belief (i.e. an emergence account favoured by authors such as Davis and Pratt). Although our data are consistent with such a non-causal account, stronger conclusions await further testing of these ideas; for instance, with different longitudinal time frames and different working memory tasks.

In summary, we argue that our data, taken with other findings, support a bidirectional relation between language and theory of mind, and are consistent with the idea that both syntax and semantics contribute to false belief understanding.

Acknowledgements

Many thanks to the Economic and Social Research Council for financial support (Award R000237071), to David Hitchin for his statistical help, and to Kate Rowlandson and Linda Murdoch for help in preparing materials. We would also like to give special thanks to the staff and children at Daisychain, Orchard Day, Royal Spa, and University of Sussex nurseries for their help and participation in this study.
References


*Received 9 July 2003; revised version received 27 January 2004*
This is a horse and this is a sheep.

Look, I put this horse down first and then this sheep.

But I’m going to say the names of these 2 things in a backwards order, so I say sheep [point] then horse [point].

Now you say them in a backwards order (pointing at each picture to help). Well done!
Picture/Verbal

Now try one, I say *teddy, ball*.
You say those backwards. Well done.

Now I’m not going to put the pictures down now, I’m just going to say the words. I want you to say what I say in a backwards order.

**TEST (all verbal)**

1. I say Glasses house You say those backwards Picture if fail
2. I say Scissors telephone You say those backwards Picture if fail
3. I say Sandwich foot You say those backwards
4. I say Plane dog pear You say those backwards
5. I say Knife boat sun You say those backwards
6. I say Apple table hat You say those backwards