METALINGUISTIC AWARENESS AND THEORY OF MIND: JUST TWO WORDS FOR THE SAME THING?

Martin Doherty
University of Sussex

Josef Perner
University of Salzburg

Three- to 5-year-old children were tested on a traditional False Belief task, in which children have to predict where a protagonist will look for an unexpectedly moved object, and a new metalinguistic task. In this task children named an item (e.g., "rabbit") and they had to monitor that another person used a synonym for naming the same item (e.g., "bunny"). Both tasks were mastered about the age of 4 years with a strong correlation between the two tasks that remained above .70 even after partialling out control measures and verbal intelligence. Moreover, younger children's difficulties with the metalinguistic task did not extend to a control task of equivalent logical structure and complexity. A simplified version of the task in which children had to produce synonyms themselves yielded very similar results. The findings confirm that metalinguistic awareness can be demonstrated around 4 years and they support the theory that the ability to understand belief relates to the development of understanding representations.

The aim of this paper is to provide evidence for a common development underlying the ability to reflect on language as a carrier of meaning (metalinguistic awareness) and the ability to understand the representational nature of mind. In

Part of the work presented here was carried out whilst Josef Perner was on sabbatical from the Laboratory of Experimental Psychology, University of Sussex at the Max-Planck Institute for Psychological Research, Munich. Experiments 1 and 2 were presented at the Biennial Meeting of the Society for Research in Child Development, Indianapolis, 1995. Experiments 3 and 4 were presented at the Biennial Meeting of the Society for Research in Child Development, New Orleans, 1993.

Direct all correspondence to: Martin Doherty, Department of Psychology, University of Stirling, Scotland FK9 4LA <mjdl@stir.ac.uk>.

Manuscript received May 2, 1996; revision accepted January 14, 1998
particular, Flavell (1988), Ferguson and Gopnik (1988), and Perner (1988, 1991) argued that success on the False Belief task is based on understanding of the representational nature of mental states. In the typical False Belief task (Wimmer & Perner, 1983) a protagonist puts an object into one of two locations. In his absence the object is unexpectedly moved to the other location. On the protagonist's return, children are asked where the protagonist will look for the object. A typical finding is that most 3-year-olds answer wrongly that he will look in the location where the object really is, while after 4 years most children answer correctly that the protagonist will look in the original, now empty location.

Perner argued that to pass the False Belief task children must understand the protagonist's belief is a mental state that represents something (the real location) as being a certain way, namely in this case, as being different from what it really is. If this characterization of the cognitive changes underlying failure and success on the False Belief task is correct, similar developments in metalinguistic awareness should be demonstrable. Children must understand language represents states of affairs (i.e., meaning) in a certain way (e.g., in terms of its formal structure).

This issue also has relevance for when children develop metalinguistic awareness. There are roughly two schools of thought: early development at the time of language acquisition around or before 2 years (the Interaction Hypothesis, e.g., Clark, 1978; Clark & Anderson, 1979; Marshall & Morton, 1978); or between 5 to 8 years (the Autonomy Hypothesis, e.g., Gombert, 1990/1992; Hakes, 1980; Tunmer & Herriman, 1984; Van Kleeck, 1982). In particular, Gombert (1990/1992) argued that behavior that has been interpreted as metalinguistic functioning before the age of 6 to 7 years is merely "epilinguistic", that is, not based on systematically represented knowledge which can be intentionally applied (Gombert, 1990/1992, p. 9).

In contrast to both schools of thought, on the basis of the theoretical analysis above, we predict that metalinguistic awareness will arise at the same time that children begin to succeed on the False Belief task, roughly at the age of 4 years. In order to test this prediction, a suitable metalinguistic task is required. Such a test must satisfy three requirements: it must

1. require understanding that language is a formal system carrying meaning;
2. require systematically represented knowledge which can be intentionally applied, in order to satisfy the concerns of Gombert and others favoring the Autonomy Hypothesis;
3. be based on simple, basic, and familiar formal aspects of language in order to be sure that children do not fail for reasons other than lacking the competence the task aims to assess.

In general, two types of evidence demonstrate understanding that language is a formal system carrying meaning. One type is understanding of the link between
grammatical form and meaning. For instance, de Villiers and de Villiers (1972, 1974) and Smith and Tager-Flusberg (1982) examined children’s ability to identify ungrammatical utterances as odd or silly. The results suggest that children begin to succeed on these tasks around the age of 4 years. For example, Smith and Tager-Flusberg (1982) found that whilst 78% of 4-year-olds successfully judged the grammaticality of sentences, only 22% of 3-year-olds could do so.

In the present study, however, we chose to examine the other type of evidence, the ability to distinguish between words (formal aspects) and what they represent (meaning). For example, many authors have asked children whether names for things can be exchanged (e.g., Piaget, 1929; Vygotsky, 1962; Rosenblum & Pinker, 1983). This method poses pragmatic problems, because swapping names is an unusual activity that is unlikely to make much sense to children. In order to avoid these problems we exploited a naturally occurring split between words and what they represent: synonymy (a similar split occurs in homonymy, and is the subject of ongoing research). Surprisingly, children’s understanding of lexical synonymy has not been addressed before in the literature (Gombert, 1990/1992).

In order to identify early competence, we are looking for a test that needs understanding of the most rudimentary and obvious aspects of formal structure and meaning relations, which are commonly within the reach of any normal adult’s reflection. The relation between form and meaning for synonyms is simple and obvious: the same meaning is expressed by two different forms. Furthermore, preschool children’s vocabularies contain a number of synonyms (as vocabulary testing and pilot work revealed) and so this relation between form and meaning is likely to be familiar to children.

The task used to assess understanding of synonyms was modeled after the, for children, highly familiar naming task. In the first two experiments, children are shown items in a picture (e.g., a rabbit) and are asked to name them with a common noun (i.e., “rabbit”). Another person (or hand puppet) then names the same item by using either the same noun as the child (“This is a rabbit”), a synonym (“This is a bunny”), or a wrong label (“This is an elephant”). The child is told that the puppet’s task is to name the item correctly (same meaning) but use a different name (different form) than the child. Hence, when asked whether the puppet had done what it was supposed to do the child should say “yes” only when the puppet had used the synonym.

One danger is that the extraneous non-linguistic demands of the task may mask metalinguistic competence, for example, memory demands, or the ability to coordinate judgments of two dimensions. To test for this possibility a structurally identical control task is used in which children have to monitor which kind of item the puppet selects. They have to ensure that the puppet selects a different item but of the same kind. If the younger children who have difficulty with the Metalinguistic task have no or only minor problems with this control task then their diffi-
culties with the Metalinguistic task cannot be attributed to its non-linguistic complexity.¹

**EXPERIMENT 1**

To test the prediction that an appropriate Metalinguistic task poses the same developmental difficulty as the False Belief task, children were tested on a version of the standard False Belief task and our Synonym Judgment task. We anticipate that most 3-year-olds will fail the False Belief and the Metalinguistic tasks but have little problem with the control task and that by 4 years most children will give correct answers on all tasks. In addition, we expect that in the transition period performance on the False Belief and Synonym Judgment tasks correlate substantially.

**Method**

**Participants.** Twenty-six children from a private preschool with largely middle class background participated in this study. Two children chose not to complete the experiment. Of the 24 children who completed the experiment, 17 were boys and 7 were girls. Their ages ranged from 3 years (3;0) to 4 years and 5 months (4;5) with a mean age of 3;9 and a standard deviation of 4 months. For the analysis of results, children were divided into two groups: a younger group (12 children from 3;0 to 3;9, mean age 3;6, SD = 2.5 months), and an older group (12 children from 3;10 to 4;5, mean age 4;0, SD = 2 months).

**Design.** Each child was tested on all three tasks: Synonym Judgment, False Belief, and Object Pointing control. The order of administration was counterbalanced in a $3 \times 3$ sequence balanced Latin square design.

The Synonym Judgment task consisted of four component trials. In two of the trials the puppet correctly used a synonym, in one trial the puppet incorrectly used the same word as the child, and in the fourth trial the puppet misnamed the item. Order of presentation was such that the first and last pair of trials each required one 'yes' judgment and one 'no' judgment from the child. Otherwise order of presentation and item named were fully counterbalanced.

The Object Pointing control consisted of four analogous component trials. In two of the trials the puppet correctly pointed to another picture showing the same kind of object; in one of the trials the puppet incorrectly pointed to the same picture as the child; in the fourth trial the puppet pointed incorrectly to a picture

---

¹In the actual test, for ease of presentation, pictures of objects were used instead of the objects themselves. This does not, of course, require any understanding of pictorial representation beyond being able to recognize the depicted situation. The task requires children to distinguish between different situations—a yellow rabbit and a brown rabbit for example—not between different pictorial representations of the same situation.
showing a different object. The order of these trials and items depicted were counterbalanced in the same way as the four Synonym Judgment trials.

**Procedure and Materials**

Each child was seen in a quiet and familiar room adjacent to the nursery area. The following three tasks were administered in the order discussed in the Design section above.

**Synonym Judgment Task**

The Synonym Judgment task consisted of three phases: vocabulary-check, modelling and test phase.

*Vocabulary-check.* In this phase children were given a vocabulary test checking on their knowledge of the synonyms used later in the actual test. It also served to alert the child to the distinctions which had to be made in the experiment. Four A4 sheets were used. Each of them had four pictures on it. Two of the pictures were experimental items used later (truck/lorry and woman/lady on two of the sheets; TV/television and coat/jacket on the other two). The other two items on each sheet were chosen from among a rabbit, a cat, an apple, a bird, and a daisy. Children were shown each sheet and asked to point to, e.g., a truck, and then to a lorry. If they hesitated they were given encouragement, and the question was repeated if they answered incorrectly. They were then told that the object has two names, lorry and truck. On the third and fourth sheets the first item to identify was not one of the experimental items, in order to prevent children from thinking that the same item was required for both questions on each sheet, and then pointing to the same item regardless of which word is used. Then the experimental item was asked about once with each synonym, as before.

*Modelling Phase.* The objective of this part of the procedure was to model the actual test procedure. A white glove puppet and three hand drawn 10 × 15 cm color pictures (showing a cup, rabbit, and settee) were used. The child was shown the picture of a cup, told that it could be called a cup or a mug, and invited to choose one of these names. The puppet’s task was “to say the other name. NOT the one that you said”. The experimenter continued:

Now, Puppet, you say the other name. [Puppet uses the same name as child, e.g., “cup”]. Is that what he should have said? [Pause for child to answer]. No, cause you said ‘cup’ didn’t you? Puppet, you say the other name. [Puppet now gives synonym, e.g., ‘mug’]. Is that what he should have said? [Wait for child’s answer]. Yes, because ‘mug’ is the other name for cup, isn’t it?

This was repeated for the other two warm-up items: rabbit/bunny, and settee/sofa. On these two trials the puppet also gave the wrong answer first but then on the second try said something completely different (e.g., elephant or banana) before finally producing the correct synonym.
Test Phase. For the actual test the modelling phase was continued with four new pictures but no feedback was given. Puppet named each item only once and then the test question: “Is that what he should have said?” was asked. The four items were always presented in fixed order as listed but the assignment of response-type was counterbalanced as explained in the Design paragraph. Depending on response-type (same, synonym, different-meaning), the puppet used one of the following words for each of the following four items:

1. coat: “coat” “jacket” “television”
2. woman: “woman” “lady” “truck”
3. truck: “truck” “lorry” “lady”
4. television: “television” “TV” “coat”

Object Pointing Control

The procedure in this task was designed to parallel the modelling and test phases of the Synonym Judgment task.

Three cards were laid in front of the child. On each card was a picture of one of the same objects used in the Synonym task. Two of the pictures were of the same object, identical but colored differently, and the other picture was of a different object (e.g., a red cup, a blue cup, and a green truck). The child was asked to point to one of the identical objects (e.g., the red cup) and Puppet’s job was “to point to the other cup, NOT the one [the child] pointed to” (i.e., the blue cup). Children were required to judge whether Puppet’s response was appropriate. The test question was: “Is that what he should have done?”

There were three trials of this procedure in the modelling phase, each with a different triplet of cards. As in the synonym modelling phase, Puppet pointed to the same object as the child and the distracter object before pointing correctly, with the experimenter providing appropriate feedback and explanation. For the test phase the procedure simply continued without feedback. Puppet made only one response for each triplet of cards. Order of presentation of items and tasks was counterbalanced as explained in the Design section.

False Belief Test

For this test a short story was acted out with two Playpeople dolls (5 cm), a marble, an opaque jar (5 cm high × 2.5 cm wide) and a box (3 cm high × 4 cm wide). In the story one of the dolls, Sally, places a marble in the box and exits. In her absence the other doll moves the marble to the jar and also leaves. Sally returns and children are asked the following questions:

Belief Question: Where will she look first for her marble?
Reality Question: Where is the marble really?
Memory Question: Where did Sally put the marble in the beginning?
Results

Figure 1 shows the percentage of children passing each of the three tasks of Experiment 1: Synonym Judgment task, Object Pointing control task, and False Belief task. In order to compare performances on the three tasks, children were counted as successful on the Synonym Judgment and Object Pointing tasks only if they answered all four trials correctly. Before considering the relationship between tasks further however, we consider performance separately.

**Vocabulary Check.** Four children made one error and one child made two errors during the vocabulary-check preceding the Synonym Judgment task. Four of these five children went on to pass the Synonym Judgment task. Thus poor performance on the Synonym Judgment task cannot be attributed to absence of the specific synonyms used from children's vocabulary, even for children who failed to identify vocabulary items.

**Synonym Judgment Task and Object Pointing Control.** For the Synonym Judgment and Object Pointing control tasks, an analysis of variance was carried out on the percent correct answers over the four trials with the two age groups as a between subjects factor and tasks (Synonym versus Object Pointing control) as
a within subjects factor. Both age-group \( F(1, 22) = 7.14, MS(\text{error}) = 0.57, p < .05 \) and task \( F(1, 22) = 5.34, MS(\text{error}) = 0.39, p < .05 \) were significant. This result confirms the trends apparent in Figure 1.

**False Belief Test.** Answers to the Belief question consisted of either indicating correctly that Sally would look in the empty box or indicating wrongly that Sally would look in the box where her marble actually was. There was the typical improvement over the particular age range tested with 42% and 75% correct responses within the two age groups, as shown in Figure 1. This improvement was statistically significant on logistic regression with age as an interval variable: \( F\text{-to-enter}(1, 22) = 8.63, p < .01; F\text{-to-remove}(1, 22) = 4.87, p < .05 \).

Performance on the control questions was good, with only one child committing an error on the reality question and one child committing an error on the memory question, both of whom also answered the Belief question wrongly.

**Comparison of Tasks**

Figure 1 shows that understanding of false belief and metalinguistic understanding emerge at about the same age, and are thus of comparable difficulty for each age group. The Figure also confirms the results of the analysis of variance, that children found the Object Pointing control much easier than the metalinguistic task.

Table 1 shows the contingencies between success on the False Belief and Synonym task. Clearly, performance on the two tasks is highly related, \( r_\phi = .83, df = 22, p < .001 \). Furthermore, partialling out performance on the Object Pointing control task, which is structurally similar to the Synonym task, does not affect the substantial and significant correlation of metalinguistic understanding with false belief: \( r = .85, df = 21, p < .001 \). This resilience of the correlation suggests that the relation between the Metalinguistic task and the False Belief task is a very specific one.

**EXPERIMENT 2**

Experiment 2 aimed to replicate and extend the findings of Experiment 1. In particular, a measure of verbal mental age was added. It is important to assess the

<table>
<thead>
<tr>
<th>False Belief</th>
<th>Synonym Judgment</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass</td>
<td>Synonym Judgment</td>
<td></td>
</tr>
<tr>
<td>Pass</td>
<td>Pass</td>
<td>13</td>
</tr>
<tr>
<td>Fail</td>
<td>Fail</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pass</td>
<td>Pass</td>
<td>13</td>
</tr>
<tr>
<td>Fail</td>
<td>Fail</td>
<td>9</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>10</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>24</td>
</tr>
</tbody>
</table>
possibility that the Metalinguistic and False Belief tasks are so strongly associated because both depend on general verbal intelligence. A wider age range was tested, and more younger children were tested to include a greater proportion who fail the False Belief task. Because only one child had failed the “different meaning” synonym trial and its object pointing analogue, this type of trial was deemed redundant. Instead, two “identical” and two synonym trials were administered to give a more consistent picture of children’s metalinguistic ability.

As an independent measure of general verbal intelligence, the short form of the BPVS (British Picture Vocabulary Test, Dunn, Dunn, Whetton, & Pintillie, 1982) was administered to each child. The short form of the BPVS has been found to be reliable in terms of internal consistency (for 3- and 4-year-old children, the split-half reliability coefficients are 0.83 and 0.84, respectively) (Dunn et al., 1982). Although statistical evidence for the validity of the BPVS is not presented, Dunn et al. argue that as the BPVS is an adaptation of the Peabody Picture Vocabulary Test, evidence supporting the validity of the PPVT (Robertson & Eisenberg, 1981) supports the validity of the BPVS. There is precedent in the metalinguistic literature for assessing linguistic development using the PPVT (e.g., Smith & Tager-Flusberg, 1982), and in the theory of mind literature for the BPVS, especially in the case of autism (e.g., Happé, 1995).

Method

Participants. Twenty-six children from a private preschool from largely middle class background participated in this study, none of whom had participated in previous studies. One child chose not to complete the experiment. Of the 25 children who completed the experiment, 13 were boys and 12 were girls. Their ages ranged from 2;5 to 4;7 with a mean age of 3;8 and a standard deviation of 8 months. For the analysis of the results, the children were divided into two groups: a younger group (12 children from 2;5 to 3;7, mean age 3;1, SD = 5 months), and an older group (13 children from 3;9 to 4;7, mean age 4;2, SD = 3.5 months).

Design and Procedure. Design and procedure were the same as in the previous experiment with the following alterations. The four component trials of the Synonym Judgment task consisted of two in which the puppet correctly used a synonym, and two in which the puppet said the same word as the child. Trials were ordered into pairs of one synonym trial and one same-word trial. Otherwise, order of presentation and item were fully counterbalanced. The Object Pointing control task consisted of four analogous trials that were administered in an analogous order to the synonym trials.

British Picture Vocabulary Scale. The short form of the BPVS was administered about a week after the original test session. Three children (aged 3;4, 3;11, and 4;0) could not be tested because they had left the nursery for the summer.
Results

Vocabulary Check. Two children failed one item of the vocabulary check. One of these children (aged 3;4) failed the Synonym and the False Belief tasks, but passed the control. The other child passed all tasks. Again, poor performance on the Synonym task cannot be attributed to absence of the specific synonyms from the child’s vocabulary.

Synonym Judgment Task and Object Pointing Control. Figure 2 shows the number of children passing each of the three tasks of Experiment 2. As before, for comparison with the False Belief task, performances on the Synonym and Object Pointing control tasks were counted as success if the child was correct on all four trials. Performance on the three tasks are first considered independently.

Analysis of variance on the percent correct answers to the four synonym judgment trials with the two age groups as a between subjects factor and tasks (Synonym versus Object Pointing control) as a within subjects factor shows that there was both a significant improvement with age ($F(1, 23) = 20.48, MS(\text{error}) = 0.67, p < .001$) and a significant difference between the Object Pointing control and the Synonym task ($F(1, 23) = 10.71, MS(\text{error}) = 0.44, p < .005$). The interaction of age group and task is also marginally significant ($F(1, 23) = 4.18, MS(\text{error}) = $
Table 2. Contingency Between Number of Children Passing/Failing Synonym Judgment and False Belief Tasks in Experiment 2

<table>
<thead>
<tr>
<th>False Belief</th>
<th>Synonym Judgment</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pass</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>Fail</td>
<td>2</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>13</td>
<td>12</td>
</tr>
</tbody>
</table>

0.44, $p = .0525$), attributable to the ceiling performance of the older children on the Object Pointing task.

**False Belief Task.** There was the typical improvement in understanding false belief with age from 8% to 85% correct responses over the two age groups. This improvement was statistically significant on logistic regression with age as an interval variable: $F$-to-enter $(1, 23) = 33.61, p < .001$; $F$-to remove $(1, 23) = 11.39, p < .005$.

Performance on the control questions was good, with only two children committing an error on the memory question (and failing the belief question), and one child committing an error on the reality question (and passing the belief question).

**Comparison of Tasks**

Figure 2 shows that false belief and metalinguistic understanding emerge at about the same age, and confirm the results of the analysis of variance that children found the Object Pointing control task much easier than the metalinguistic task.

Table 2 shows the contingencies between success on the False Belief and Synonym tasks. As in Experiment 1, performance on the two tasks is highly related, $r_0 = .82, df = 20, p < .001$. This substantial, highly significant correlation of metalinguistic understanding with false belief remains so even after performance on the BPVS (measuring verbal mental age) and on the Object Pointing control have been partialled out, $r = .70, df = 18, p < .001$. Thus the correlation cannot be attributed to a general increase in verbal mental age, or to the general information processing demands of the Synonym task, some of which may be common to the False Belief task.

**EXPERIMENT 3**

Experiment 2 replicates the essential findings of Experiment 1, that false belief understanding and the judgment of synonymy are strongly related. The following two experiments aim to provide a broader picture of children’s understanding of synonymy in a new task that simplifies the demands made on children. The judgment method employed in the first two experiments is a formal “school-like” sit
uation that children in preschool may not find familiar, and relies on yes-no questions which children may have difficulty with. A more direct, simpler method is to give children one synonym and ask them to produce the other (essentially casting children in the role of the puppet in the previous two experiments). Success on this task without metalinguistic understanding would be possible half the time if the experimenter provides the less common label (e.g., "rabbit") and children, following their natural labelling tendency, then produce the synonym (e.g., "bunny"). To guard against this possibility, the criterion for success is producing the synonym in both directions for a pair (e.g., rabbit-bunny and bunny-rabbit) tested on separate occasions. This task also allows for a range of success (in terms of number of pairs produced), unlike the judgment task in which success was only granted if children were correct on all four trials. Thus the new task has greater potential to show early metalinguistic understanding.

Method

Participants. Forty-four children participated in this study, none of whom had participated in previous studies. Six children chose not to complete the experiment. Of the 38 remaining children, 8 were from a predominantly working class play group in Brighton (mean 3;8, $SD = 7.6$ months), 21 were from an upper working class nursery in Sheffield (mean age 4;3, $SD = 4.0$ months), and 9 were from a middle class nursery in Sheffield (mean age 3;9, $SD = 4.9$ months). Seventeen of the children were girls and 21 were boys. For the analysis of the results, children were divided into two groups: a younger group (20 children from 2;9 to 4;1 months, mean age 3;8 $SD = 4.7$ months), and an older group (18 children from 4;2 to 4;11, mean age 4;5, $SD = 2.5$ months).

Design, Procedure and Materials. Each child was seen in a familiar quiet room adjacent to the nursery. The Synonym Production and False Belief tasks were administered in counterbalanced order. Half the children received the BPVS at the start of the session and half received it at the end of the session. Only the Synonym Production Task differed from the previous experiments.

Synonym Production Task

The Synonym Production Task consisted of three phases: vocabulary-check, modelling phase, and test phase.

Vocabulary Check. The vocabulary check was similar to the previous experiments, except that children were not explicitly alerted to the fact that synonyms were being used. Pictures of items occurred on two of the four sheets; on the first presentation of a picture one synonym was used, and on the second presentation a few items later the other synonym was used.

Modelling and Experimental Phases. The procedure was similar to the judgment procedure of Experiments 1 and 2, but the child had to produce synonyms to
help Puppet, rather than judge his output. Colored pictures were used as before. A
Playpeople doll (5 cm) “provided” one synonym and children had to help Puppet
produce the other synonym. The following prompts were used if the child did not
answer:

Prompt 1: Can you think of another way of saying Jumper?
and then if necessary
Prompt 2: What’s another word for Jumper?

If children repeated the same word as the doll they were told, “But Tony called it
a rabbit. Puppet wants to say something different”. If children misnamed the
object they were told, “But it isn’t really a frog is it? Puppet wants to say what it
really is, but in a different way”. For the three modelling items (rabbit/bunny, cat/pussy, settee/sofa) the experimenter provided the synonym. In the experimental
phase the child had to provide the synonym. If they did not do so following the
prompts they were reassured with phrases such as “Never mind” or “That’s a hard
one isn’t it?” and the next item presented.

All five experimental items were presented twice in fixed order. On the second
presentation the doll used the other synonym. The order of items and first syn-
onym/second synonym were as follows:

1. Jumper/Sweater
2. Lady/Woman
3. Cup/Mug
4. TV/Television
5. Coat/Jacket

Results

Vocabulary-Check. The majority of children knew all of the synonyms used.
Fourteen of the 38 children failed to identify one or more of the experimental
items: 8 children failed 1 vocabulary item, 3 children failed 2, and 3 children
failed 3 vocabulary items. No child failed to indicate both items of a synonym
pair. Failures were not evenly spread: 9 children failed Sweater, 7 failed mug, 3
failed jacket, and one child failed to identify woman, mug and TV. Vocabulary
failures will be partialled out from the later analysis relating performance on the
False Belief tasks to performance on the Synonym Production Task.

Synonym Pair Production. Figure 3 shows the number of children producing
0-5 synonym pairs. Fifteen children produced no pairs at all with the remaining 23
children producing a mean of 3.6 pairs each ($SD = 1.2$ pairs). Performance tended
towards either general success or failure, with 30 of 38 children producing either
no pairs or 4 or 5 pairs.
Figure 3. Number of children producing 0-5 synonym pairs and showing false belief understanding in Experiment 3.

Overall, the mean number of synonym pairs produced was 2.2. The mean number of synonym pairs for which children correctly identified both synonyms in the vocabulary check was 4.3. This can be seen as a measure of the number of pairs children could potentially produce if they had metalinguistic awareness. To assess whether this differed from the number of pairs actually produced, an analysis of variance was carried out on the mean number of pairs with the two age groups as a between subjects factor and task (production versus vocabulary) as a within subjects factor. Task was significant ($F(1, 36) = 47.53, MS$ (error) = 1.83, $p < .001$), but age group was not ($F(1, 36) = 0.00, MS$ (error) = 3.38). This indicates that children of both age groups' difficulty in producing synonyms was not due to their poor vocabulary.

The lack of improvement between the age groups seems to be largely an effect of school. Children from the working class nursery school, who made up the bulk of the older children (mean age 4;3) produced a mean number of 1.5 synonym pairs, far less than the predominantly younger children from the working class playgroup (mean age 3;8, pairs = 2.3) and particularly the middle class children (mean age 3;9, pairs = 3.7).
**False Belief Test.** Fifty-three percent of children were successful on the False Belief task. Although performance improves with age with 40% of the younger group compared to 67% of the older group passing, this is not significant on logistic regression with age as an interval variable \( F\text{-to-enter (1, 36)} = 2.06, F\text{-to-remove (1, 35)} = 1.83 \). Four of the 38 children failed the False Belief memory question, all of whom failed the belief question.

**Comparison of Tasks**

Figure 3 also compares performance on the False Belief task with number of synonyms produced. Clearly synonym production and false belief performance are strongly associated, \( r = .71, df = 36, p < .001 \). This substantial and significant correlation of metalinguistic understanding with false belief remains substantial and highly significant even after performance on the BPVS and the number of failures in the vocabulary check have been partialled out, \( r = .66, df = 34, p < .001 \). The resilience of this correlation shows that, like the ability to judge synonyms, the relationship between the ability to produce synonyms and the ability to pass False Belief tasks is a very specific one.

**Experiment 4**

The final experiment aimed to replicate and extend the findings of Experiment 3. A task analogous to the Synonym task was added to control for possible unwillingness or inability to think of a different answer to the experimenter. Children were required to misname the objects by pretending that they were other objects. Thus they were required to use another name for the object but did not have to consider the relationship between the word and its referent. An additional False Belief task was added to give a more consistent picture of children’s false belief understanding. In the new False Belief task, modelled on one used by Perner, Leekam, and Wimmer (1987), the protagonist expects the object to be moved when in fact it is not. This guards against any tendency children might have to choose the initial location of the object—and thus be scored as correct—without considering belief.

**Method**

**Participants.** Thirty-seven children participated in this study, none of whom had participated in previous studies. Six children chose not to complete the experiment. The 31 remaining children came from five middle class nursery schools and playgroups in Brighton. Nineteen of the children were girls, and 12 were boys. For the analysis of the results, children were divided into two groups: a younger group (15 children from 2;10 to 3;9, mean age 3;3, \( SD = 3.6 \) months), and an older group (16 children from 3;10 to 4;7, mean age 4;2, \( SD = 3.0 \)).
Design. Each child was tested on all five tasks: Synonym Production, two False Belief tasks (Unexpected Change and Unexpected No-change), Pretend Control Task, and BPVS. Children were tested over two sessions a week apart; any one session included only one of the False Belief tasks and either the Synonym or the Pretend task, in counterbalanced order.

Procedure and Materials

Synonym Production Task. The Synonym Production Task was the same as in Experiment 3, apart from changes in items used to reduce the number of vocabulary failures of Experiment 3. Also a different, more sober looking puppet was used to prevent confusion between the Synonym and Pretend Production tasks. He was referred to as Sensible Puppet.

Pretend Control Task. The Pretend Task was designed to parallel the Synonym Production Task in form. Instead of a vocabulary check, children were shown the same set of A4 sheets and asked, “Which one could we pretend is a crocodile?” After children had selected one, the experimenter continued, “And which one could we pretend is a tree?” and so on, asking one such question for each of the items used in the experiment.

For the modelling and experimental phases the same set of index card pictures and the doll used in the Synonym task were used in the same order. The puppet used in the previous experiments was introduced as Silly Puppet who likes to say silly things, and pretend that one thing is another thing. The procedure was modelled by the experimenter for the first three items, with feedback and explanation, then continued without feedback for the experimental items. Items were only presented once.

Test Question: What could puppet call it?
Prompt 1: What could puppet pretend it is?
Prompt 2: What’s a different thing we could pretend it is?

False Belief Tasks

One task (Unexpected Change) was the same as used in previous experiments. The materials for the new (Unexpected No-Change) task were the same. However, in the story, Tony leaves his ball in the jar, and asks Sally to put it into the box when she has finished playing with it. She forgets, and thus Tony falsely believes the ball to be in the box when he returns:

Belief Question: Where will he look first for his marble?
Reality Question: Where is the marble really?
Memory Question: Where did Tony ask Sally to put the marble?
Results

**Vocabulary Check.** The modifications in test items appear to have been largely successful. Only six children failed to identify one of the test items, and the item was “mug” in each case.

**Synonym Production.** Figure 4 shows the number of children producing 0-5 synonym pairs. Six children produced no pairs at all. The remaining 25 children produced an average of 2.8 pairs each ($SD = 1.2$ pairs). Performance is relatively evenly spread compared to performance in Experiment 3 (see Figure 3).

Children produced on average 2.3 synonym pairs ($SD = 1.6$), far less than the number of pairs for which children identified both synonyms in the vocabulary check, 4.9 ($SD = 0.4$). This was confirmed by an analysis of variance on the mean number of pairs with the two age groups as a between subjects factor and task (production versus vocabulary) as a within subjects factor. Task was significant ($F(1, 29) = 107.22$, $MS$ (error) = 1.00, $p < .001$), age was marginally significant ($F(1, 29) = 3.71$, $MS$ (error) = 1.23, $p = .064$), and the interaction between age and task was significant ($F(1, 29) = 4.46$, $MS$ (error) = 1.00, $p < .05$) (which is accounted for by the ceiling performance on the vocabulary test compared to an increase in production performance with age).
**Pretend Task.** Success on the Pretend task was high, with 24 out of 31 children producing a pretend alternative for all five items. Children produced a mean of 4.2 pretend items ($SD = 1.7$), much greater than the mean of 2.3 synonym pairs.

To assess whether children produced significantly more pretend items than synonym pairs, an analysis of variance was carried out with the two age groups as a between subjects factor, and task (pretend versus synonym pairs) as the within subjects factor. Both age and task were significant: Age, $F(1, 29) = 4.41$, $MS$ (error) $= 3.20$, $p < .05$; task, $F(1, 29) = 30.65$, $MS$ (error) $= 1.78$, $p < .001$. Thus performance on the Pretend control was substantially and significantly better than performance on the Synonym task which shows that children’s problem with the Synonym task is not a reluctance to freely generate some different answer. Their difficulty must have to do with generating an alternative description of an item.

**False Belief Tests.** Fifty-five percent of children passed the Unexpected Change task and 48% of children passed the Unexpected No-Change task. Twenty-seven of the 31 children passed or failed both tasks, and the two tasks were highly correlated ($r_g = .75$, $\chi^2(1, N = 31) = 14.5$, $p < .001$) and did not differ significantly (Binomial: $N = 4$, $k = 1$, $p = .312$). Because the performances on the two belief questions were so highly related, scores were added to yield a compound false belief score from 0 (fail both tasks) to 2 (pass both tasks).

Although there was improvement across the age range for both False Belief tasks, the improvement was not significant on logistic regression with age as an interval variable and only compound false belief scores of 2 considered a pass: $F$-to-enter $(1, 29) = 1.54$, $p = .22$; $F$-to-remove $(1, 29)$, $p = .60$.

One child failed the memory question on the Unexpected Change task, but passed the test question. Seven children failed the slightly more complex memory question on the Unexpected No-Change task, only one of whom passed the test question.

**Comparison of Tasks**

Figure 4 also shows performance on the False Belief tasks in comparison to the number of synonyms produced. Again, synonym production and false belief performance are strongly associated, $r = .65$, $df = 29$, $p < .001$. This substantial, significant correlation of metalinguistic understanding with false belief remains strong even after performance on the BPVS and the number of pretend items produced have been partialled out, $r = .60$, $df = 27$, $p < .001$.

**SUMMARY AND COMBINED ANALYSIS**

Experiments 3 and 4 reinforce the findings of Experiments 1 and 2. Understanding of synonymy, whether measured through production or judgment, is highly
associated with false belief understanding, even after the effects of verbal mental age (Experiments 2, 3, and 4) and performance on control tasks (Experiments 1, 2, and 4) are controlled for.

An overall picture of the relation between understanding of synonymy and false belief can be obtained by combining the results from all four experiments. On the judgment task, children only succeed if they give the correct answer on all 4 trials, whereas the production task allows a range of success from 0 pairs produced to 5. In order to combine results from the two tasks then, performance on the production task also has to be classified as success or failure according to some criterion. A certain amount of leniency is desirable to allow for lapses in attention, vocabulary gaps and word-finding difficulties, but one out of five pairs correct may be overly lenient. As a simple compromise we set the criterion at three out of five pairs correct.2

Table 3 compares success on the Synonym tasks (both judgment, and production employing the criterion just discussed) and the False Belief task used in all experiments (i.e., Unexpected Change in Experiment 4). As can be seen, the correlation found between Synonym and False Belief task performances in each of the experiments is also strongly visible when the results are combined. Roughly equal numbers of children passed and failed each task, and of 118 children who were tested, 100 passed both or failed both the Synonym and the False Belief task. The correlation between the two types of task was $r_{phi} = .70$, $df = 116$, $p < .001$. (Unfortunately, since the BPVS was not administered in all four experiments, verbal mental age scores cannot be partialed out).

Thus the combined results of all four experiments reinforce the conclusions from each individual experiment: the relation between false belief understanding and metalinguistic awareness (at least the particular skill investigated) is strong, and specific.

---

2 In practice, the precise level of this criterion does not substantially alter the relation between false belief and Synonym task performances: overall correlations between false belief and synonym understanding for the different possible production criteria were as follows: 1/5 pairs, $r = .67$, 2/5, $r = .71$, 3/5, $r = .70$, 4/5, $r = .63$, 5/5, $r = .44$, $p < .0001$ in each case.
GENERAL DISCUSSION

The results from all four experiments converge on the following three findings:

1. an aspect of metalinguistic awareness can be experimentally demonstrated around the age of 4 years;
2. younger children's difficulty with the particular metalinguistic awareness tasks cannot easily be attributed to the complexity of the task because these children had little difficulty with the Object Pointing or Pretend Production control tasks, and;
3. there is a close relation between the Metalinguistic Awareness tasks and the False Belief task. This relation is very specific beyond a common correlation with verbal intelligence (BPVS, Experiments 2, 3, and 4), and various task comprehension measures (Object Pointing Control in Experiments 1 and 2 or Pretend Control in Experiment 4).

Theory of Mind

Representational Theory of Mind. The close association between the False Belief task and the metalinguistic Synonym tasks was predicted from the theory that success on the False Belief tasks underlies an emergent understanding that one can conceive of a state of affairs as being a certain way (Perner, 1991, 1995), that is, the mistaken protagonist conceives of the object's real location as being different from what it really is. This distinction is also required for understanding representations which represent things or state of affairs as being a certain way (Goodman, 1976; Perner, 1991). To solve the Synonym task, the child has to realize that what something is can be described (re-presented) in different ways (e.g., as "a bunny" or as "a rabbit"). Hence there is a close relation between the False Belief task and metalinguistic awareness, as assessed by our Synonym task.

This does not mean that theory of mind and metalinguistic awareness (of language as a formal system with meaning) are equivalent; they are not in fact "just two words for the same thing", but both are based on a common conceptual basis. According to this interpretation, performance on the False Belief task should not only correlate specifically with the Synonym task but also with other metalinguistic awareness tasks that require an understanding of the interrelation between formal aspects of language and meaning. One such task would be the one where children are required to distinguish grammatical from ungrammatical strings (de Villiers & de Villiers, 1972, 1974; Smith & Tager-Flusberg, 1982), provided any confounding of grammaticality and oddness of meaning can be avoided. Another task would be a homonyms task where the child's task is to point to the alternative interpretation of the word for the item selected by the puppet. For instance, if the puppet points to bat the mammal in response to "Point to the bat!" then the child should point to the cricket bat but not again to the animal. It remains to be shown whether these tasks will also correlate with the False Belief task.
Apart from awaiting further empirical confirmation of the metarepresentation theory it also needs to be evaluated against potential alternative explanations of the existing data.

The Linguistic Complementation Hypothesis. Although the relation between the Synonym tasks and the False Belief task was largely independent of verbal mental age as assessed by the BPVS, some other linguistic development might be invoked to explain the relation. De Villiers (1995) has suggested that the development of false belief understanding is a result of a syntactic development in the understanding of complementation. As this theory has not been developed in any detail, we try to look here at a favorable interpretation for explaining our result. Mental state verbs and verbs of communication take complements, embedded clauses such as “it is in the box” in the sentence “Sally thinks it is in the box”. The grammar of complementation makes clear when replacement by co-referential expressions in the complement affects truth or not, understanding of which is a necessary ingredient for understanding false belief. Hence, the acquisition of this grammar provides a necessary mental requisite for thinking about belief. This grammar may, of course, also be necessary for understanding our metalinguistic task, namely which word one can use to replace the original one and leave the meaning or truth unchanged.

Although this theory may have potential in explaining our finding, there remain important open questions. In particular, it needs to be explained why substitution of co-referentials is used too liberally in the False Belief task (i.e., substituting the description of the desired object’s real location for its believed location), whereas it is used too conservatively in the Synonyms task (i.e., synonyms are not considered suitable substitutes for each other). Moreover, it is not clear at this stage, whether the grammar for complements does not need something like the concept of representation in order to fulfill the task of regulating replaceability of co-referential expressions.

Non-Belief Specific Difficulties. Critics of the tests to assess children’s theory of mind claim that the child understands belief but is prevented from demonstrating this understanding by information processing demands or misleading aspects of the task. These accounts have a common problem in accounting for the specific association between the Synonym tasks and the False Belief task. Because at face value the Synonym and False Belief tasks are very different, it is implausible that they share the same information processing difficulties (e.g., Fodor, 1992; Leslie, 1994), or cause comparable misunderstandings between children and experimenter (Siegal & Beattie, 1991).

Leslie (1994) suggests that difficulties with False Belief tasks arise because they require reconstructing the content of a belief purely on the basis of the protagonist’s exposure history, and belief and non-mental representation tasks pose difficulties to the extent that they require such reconstruction. In contrast to the
False Belief task, however, our Synonym tasks require no reconstruction—the content can be read off the picture presented to children and from the utterances of the experimenter, child, or puppet. Thus Leslie’s theory does not provide an alternative account for the close association between False Belief task and Synonym tasks.

Mutual Exclusivity

An inability to suppress the salient reality (e.g., the real location of the marble) has been proposed as a cause of failure in the False Belief task (Mitchell, 1994; Russell, Mauthner, Sharpe, & Tidswell, 1991; Zaitchik, 1991). A parallel assumption for language would be that only one label ‘really’ applies to an object, and this could explain bad performance in the judgment and production tasks. This is equivalent to a principle proposed by a number of authors whereby children tend to map words onto mutually exclusive categories (Markman, 1989; Clark, 1973; Slobin, 1978). Experimental evidence shows that 2- to 6-year-old children will map novel words onto novel objects more frequently than objects for which they already know a name (e.g., Markman & Wachtel, 1988; Merriman & Bowman, 1989). There is also anecdotal evidence of young children rejecting novel labels for objects with known labels (Macnamara, 1982; Mervis, 1987). This implies a resistance to learning synonyms and might be thought to account for our findings. For example, Flavell (1988) has suggested that theory of mind tasks and mutual exclusivity tasks both call for the understanding that objects can be represented in more than one way.

Flavell’s explanation is similar to our own, in that developing understanding of representation allows children both to pass the False Belief task and to acknowledge that one object may have more than one name. More specifically, however, we claim that prior to passing the Synonym tasks children do not understand the representational relation between words and their referents. Thus we would predict that if younger children have a mutual exclusivity bias it is not metalinguistic, that is, not in the form of a belief that one object can only have one label.

Support for this position comes from the vocabulary check used in Experiments 1 and 2. Almost all children were willing on the same occasion to point to a single object when identified under each synonym. If younger preschoolers had an explicit one object-one label assumption, they would presumably have rejected one of these labels.

Furthermore, such an assumption would produce characteristic patterns of errors. In the judgment tasks, assuming that only one label is correct would lead children to endorse puppet’s repetition of the term they themselves provided whilst rejecting puppet’s production of synonyms and misnomers. This was clearly not the case, because only 1 child showed this pattern in Experiment 1, and no children did so in Experiment 2.

In the production task, it would lead children to produce the ‘correct’ term regardless of the term produced by the experimenter. Thus children would have
produced a large number of single synonyms but no pairs. In Experiments 3 and 4 however, most children produced either no single synonyms or just one. Furthermore, if a mutual exclusivity bias accounted for failure on both false belief and Synonym tasks there should be a relation between the number of single synonyms produced and failure on the False Belief task. There was little evidence of this: those who failed this task produced a mean of 1.9 single synonyms compared to a mean of 1.5 single synonyms produced by those who passed.

Finally, it is clear that failure to learn synonyms does not account for our data, because the vocabulary checks in all four experiments show that the synonym pairs used were already in the vocabularies of even the younger children. This highlights an important difference between our study and work on the mutual exclusivity bias. Studies of the bias generally concern children's willingness to learn new words, whereas the work presented here concerns children's willingness to interchange known words when those words violate mutual exclusivity. This crucial difference helps explain why different phenomena occur in the two types of study.

**Metalinguistic Development**

The fact that children around 4 years succeed on our metalinguistic awareness task has relevance for the discussion about when children develop metalinguistic awareness. Our data are clearly inconsistent with the view that metalinguistic awareness develops between 5 to 8 years (e.g., Gombert, 1992; Tunmer & Hermann, 1994).

In defence of this view Gombert makes a distinction between “skills observed in spontaneous behavior ... and ... abilities which are based on systematically represented knowledge and can be intentionally applied” (Gombert, 1992, p. 9). Clearly, evidence for metalinguistic awareness from our task is not based on spontaneous behavior. Rather, children are asked to monitor—hence systematically represent and intentionally apply relevant knowledge to—the formal and semantic aspects of linguistic expressions. Consequently, the general success of our 4-year-old children shows that children at this age have genuine metalinguistic awareness—even according to Gombert’s very strict criteria.

The successful demonstration of relatively early competence we attribute to the fact that our task requires understanding of only the most basic formal aspects of language (difference and sameness of words), that our task is a very familiar context even for the youngest children, and that the task assesses an aspect of metalinguistic awareness which is in the repertoire of every normally developed adult. In contrast, most traditional investigations of metalinguistic competence relied on the understanding of relatively complex linguistic features (e.g., grammatical well-formedness) that are not typically reflected upon in everyday life outside a school context.

Our study also shows that children younger than 4 years find our metalinguistic task difficult. Their comparatively good performance in the structurally identical Object Pointing or Pretend Production control tasks shows that their problems are not due to the complexity of the metalinguistic task. This, in turn, suggests that
they lack the metalinguistic competence required by our task (i.e., understanding of language as a formal system carrying meaning).

Of course, from this it does not follow that younger children could not succeed on other metalinguistic tasks. Some tasks commonly included in the metalinguistic literature do not require attention to both the meaning and formal aspects of language, and therefore should have no specific relation to either the False Belief task or our Synonym tasks. For example, Smith and Tager-Flusberg (1982) had children distinguish meaningless speech sounds (e.g., [ne], [ba], etc.) from nonspeech sounds (pop, hum, etc.). This and similar tasks that examine children’s phonemic awareness typically require distinctions purely on the level of sounds without concern that the speech sounds are speech sounds because they are constituents of a meaning carrying system. Similarly, with the skills observed in spontaneous behavior (spontaneous self correction, Clark, 1978; play with rhymes: Slobin, 1978; Leopold, 1949), it is not clear whether children pay attention to both the formal and semantic aspects of language.

Arguably, behaviors that do not require attention to both formal and semantic aspects of language do not conform to the common intuition behind definitions of metalinguistic awareness. There is no single accepted definition; definitions vary from the very strict, which deny metalinguistic status to behaviors that cannot be produced at the experimenter’s demand (e.g., Gombert, 1992), to the lenient, such as Bowey’s (1988) definition as “the ability to reflect on and manipulate the structural features of language” (p. 3). Common to all definitions (sometimes implicitly) is the notion that metalinguistic awareness is awareness of language as language. That is, children must not only be able to attend to the form of language, but must also be aware of its function, to carry meaning.

It is this central feature of metalinguistic awareness our Synonym tasks were designed to test, and the data suggest that it arises about the age of 4 years. Further research with different tasks will be needed to confirm that this is the case, but as yet there is no clear evidence for this kind of understanding before the age of about 4 years apart from some anecdotes. For instance, Slobin (1978) reported that his daughter at 3 years and 2 months made explicit remarks about the synonymy of two words. Such anecdotes, however, may be the rare productions of linguistically precocious offspring and are quite compatible with our finding that this ability does not develop in the general population until around the age of 4 years. In fact, such precocious anecdotes are quite in keeping with the spread of ages at which understanding of false belief is observed (Perner et al., 1987; Perner, Ruffman, & Leekam, 1994), with which performance on our metalinguistic task is so strongly associated.

**Conclusion**

A developing understanding of the representational nature of language around the age of 4 years was predicted from the theory that success on the False Belief task
at that age depends on an understanding of the representational nature of mental states. All four experiments supported this prediction, and showed a strong and specific relation between the False Belief task and our novel Synonym task. From this we conclude that the central feature of metalinguistic awareness, understanding language as language, develops around this age. Similar relation should be demonstrable with other suitable metalinguistic tasks and the False Belief task, since both types of task are based on a common conceptual understanding that one and the same state of affairs can be conceived in different ways. Children should begin to show genuine metalinguistic awareness from this age.

REFERENCES


